Mathematical Biosciences and Engineering

MBE, 18(3): 2150–2181.
DOI: 10.3934/mbe.2021109
Received: 23 December 2021
Accepted: 18 February 2021
Published: 03 March 2021

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

Research article

Progression and transmission of HIV (PATH 4.0)—A new agent-based evolving network simulation for modeling HIV transmission clusters

Sonza Singh[1], Anne Marie France[2], Yao-Hsuan Chen[2], Paul G. Farnham[2] Alexandra M. Oster[2], and Chaitra Gopalappa[1,*]

[1] University of Massachusetts Amherst, Amherst, MA, United States
[2] Division of HIV/AIDS Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA, United States

*** Correspondence:** E-mail: chaitrag@umass.edu; Tel: +1-413-545-2306.

Abstract: We present the Progression and Transmission of HIV (PATH 4.0), a simulation tool for analyses of cluster detection and intervention strategies. Molecular clusters are groups of HIV infections that are genetically similar, indicating rapid HIV transmission where HIV prevention resources are needed to improve health outcomes and prevent new infections. PATH 4.0 was constructed using a newly developed *agent-based evolving network modeling* (ABENM) technique and *evolving contact network algorithm* (ECNA) for generating scale-free networks. ABENM and ECNA were developed to facilitate simulation of transmission networks for low-prevalence diseases, such as HIV, which creates computational challenges for current network simulation techniques. Simulating transmission networks is essential for studying network dynamics, including clusters. We validated PATH 4.0 by comparing simulated projections of HIV diagnoses with estimates from the National HIV Surveillance System (NHSS) for 2010–2017. We also applied a cluster generation algorithm to PATH 4.0 to estimate cluster features, including the distribution of persons with diagnosed HIV infection by cluster status and size and the size distribution of clusters. Simulated features matched well with NHSS estimates, which used molecular methods to detect clusters among HIV nucleotide sequences of persons with HIV diagnosed during 2015–2017. Cluster detection and response is a component of the U.S. *Ending the HIV Epidemic* strategy. While surveillance is critical for detecting clusters, a model in conjunction with surveillance can allow us to refine cluster detection methods, understand factors associated with cluster growth, and assess interventions to inform effective response strategies. As surveillance data are only available for cases that are diagnosed and reported, a model is a critical tool to understand the true size of clusters and assess key questions, such as the relative contributions of clusters to onward transmissions. We believe PATH 4.0 is the first

modeling tool available to assess cluster detection and response at the national-level and could help inform the national strategic plan.

Keywords: agent-based simulation; network modeling; HIV modeling; network cluster analyses; infectious disease modeling

1. Introduction

In the United States, an estimated 1.15 to 1.2 million people above the age of 13 years were living with a diagnosed or undiagnosed human immunodeficiency virus (HIV) infection [1] and 37,515 persons received diagnoses of HIV infection in 2018 [2]. In addition to the disease burden, the economic burden of HIV is high, with estimated lifetime treatment costs of $250,000 to $400,000 per person [3]. However, new advancements in HIV testing, treatment for HIV-infected persons, and pre-exposure prophylaxis (PrEP) for uninfected persons at elevated risk of infection can prevent transmission. Testing allows for early diagnosis and thus treatment initiation, consistent HIV treatment helps achieve and maintain viral suppression and can result in effectively no risk of sexual transmission, and PrEP provides a preventative option for uninfected persons at an elevated risk of HIV acquisition [4,5]. The challenge is to identify the most effective strategy for allocation of resources to at-risk populations to ensure HIV prevention.

Simulation models have played a key role in identifying populations at-risk for HIV [6] and evaluating population-specific interventions to inform implementation of focused intervention programs [7,8]. One approach has been to stratify the national population into different groups based on demographic and behavioral factors, evaluate the risk of transmission for each group [9–12], and evaluate the impact of alternative intervention strategies specific to each group [13–16]. Common population stratifications have included combinations of transmission risk group (heterosexuals, men who have sex with men (MSM), and persons who inject drugs (PWID)), age group, race/ethnicity, and HIV care continuum stage of infected persons [17]. US National HIV Surveillance System (NHSS) [4] data indicate significant differences in HIV diagnosis, care, and treatment across these population groups, suggestive of differences in risk across these groups and thus the need for more focused interventions.

Recent studies using HIV nucleotide sequence data, routinely collected as part of NHSS, have indicated that molecular sequence analysis can identify networks among which HIV transmission is occurring rapidly (i.e., substantially higher than average transmission rates) [18,19]. These findings identify a new method for identifying networks of persons at increased risk of HIV infection and could allow for developing a more focused and effective approach to intervention by ensuring that prevention resources effectively reach those in need. Molecular sequence analysis identifies clusters, or groups of HIV infections that are genetically very similar. Because HIV evolves rapidly, very similar HIV nucleotide sequences indicate that HIV transmission is occurring rapidly in a common network [19]. While an average of 3.5 HIV transmission events are estimated to occur per 100 persons living with HIV per year in the United States, analysis of nucleotide sequence data collected through the NHSS identified rapidly growing clusters with an average of 44 transmission events per 100 persons per year [19]. Clusters are thus indicative of where transmission of HIV is rapidly occurring and where prevention resources would be most useful in curbing new infections.

The response to HIV clusters and outbreaks is a key part of one of the four strategic pillars of the U.S. Ending the HIV Epidemic initiative [20]. The goal of the strategic plan is to reduce new infections by 75% by 2025 and 90% by 2030. The initiative includes 4 strategic pillars to reach this goal, namely: 1) to diagnose all PWH as early as possible, 2) to treat the infection rapidly and effectively to achieve sustained viral suppression [5], 3) to prevent new HIV transmissions by using proven interventions, including pre-exposure prophylaxis [21–25] and syringe services programs [25–30], and 4) to respond quickly to potential HIV outbreaks to get needed prevention and treatment services to people who need them. Cluster detection is a key to identify potential outbreaks rapidly to guide response efforts, including implementing interventions for early diagnosis and treatment of HIV-infected persons and prevention options for those in associated networks who are at high risk of infection.

A simulation model that accurately replicates HIV transmission networks in the United States and its cluster dynamics is a key tool to evaluate and guide cluster-based prevention strategies. We present a new version of the Progression and Transmission of HIV (PATH 4.0) simulation model, constructed using a newly developed agent-based evolving network modeling (ABENM) simulation technique, and a new network generation algorithm, evolving contact network algorithm (ECNA) [31]. The earlier version of PATH [6] simulated only infected persons as agents and modeled all characteristics including demographic, sexual behavior, and partnerships as features of the agents without explicitly generating the contact network. PATH 4.0 uses the disease progression module from the earlier version of PATH [6], and reconstructs the full computational structure using ABENM and ECNA (creating the hybrid compartmental and ABNM structure), adding new modules for simulating the functionalities related to partnership formations, demographics, and sexual behavior and maintain the overall network statistics. The new simulation technique and new modules are essential for accurately simulating HIV transmission networks in the United States, and thus, for the analysis of cluster detection and response strategies. Current simulation techniques, and thus previous versions of PATH that were built using these methods [6], are infeasible to use for this application, for reasons which we will discuss in the next few paragraphs.

1.1. Current simulation methods in the HIV literature

There are two major types of dynamic simulation techniques used in national-level HIV modeling: compartmental modeling (also known as differential equations modeling) and agent-based network modeling (ABNM) [32]. Compartmental modeling splits the population into groups (or compartments) that represent the different states of a disease, e.g., susceptible, infected, and removed, and use a system of differential equations to simulate the rates of change for transitioning between these compartments. These models assume random mixing between people in a group, making them not suitable for representation of sexual and needle sharing contact networks that are known to follow a non-random scale-free network structure [33], where the distribution of the number of contacts per person follows a power-law distribution [34]. ABNM simulates infected and susceptible persons at the individual-level and the interactions leading to HIV transmissions through a contact network structure [35], which is ideal for modeling non-random network structures [34]. However, due to low prevalence of HIV in the United States, it is computationally challenging to apply ABNM to simulate HIV transmissions at the national level.

Taking data from the year 2015 as an example, the computational challenge can be described as follows. The estimated overall prevalence of HIV among persons aged 13 years or older in the United

States averaged about 419 per 100,000 persons nationally and ranged from an average of 64 to 852 per 100,000 persons by area of residence except for District of Columbia, which had an estimated average of 3,018 per 100,000 persons [36]. Clusters identified through molecular analyses of recently diagnosed infections (recent defined as cases diagnosed over the past 3-year period preceding the year of analyses, who in 2015 constituted about 10% of all PWH), range from 2 to 49 persons per cluster [19]. These cluster sizes shed light on the size of the underlying contact network structures to simulate. As ABNM are scaled versions of the population, simulating a population of 100,000 persons representative of the U.S. population would result in 419 HIV-infected persons, and 42 persons with recent diagnosis. These small samples of HIV-infected persons are insufficient to generate the clusters, in numbers and sizes, observed in molecular data. Further, the small samples create challenges in modeling heterogeneity by age, gender, and transmission risk-group, which are key categories for HIV because of differences in a disease's epidemiologic features [36–38]. As there is considerable contact mixing between these groups [39], they cannot be modeled separately. As ABNMs track interactions among N individuals, where N is the population size in the simulation, computation times are in the order of $O(N^2)$, and thus, increasing the value of N is also not a suitable solution. These issues make ABNM insufficient to use for low prevalence diseases such as HIV.

1.2. A new agent-based evolving network modeling technique for HIV

To overcome the challenges with the application of current simulation techniques for HIV cluster detection, we developed the PATH 4.0 model using a new stochastic simulation technique ABENM. When using the ABNM simulation technique, the full contact network, i.e., the network of all infected and susceptible persons, is initially generated using a network generation algorithm [35], and disease transmissions are simulated on this network over time. As sexual contact networks are known to follow scale-free network structures [40,41], a preferential attachment algorithm would be first used for generation of scale-free networks [35] in ABNM. Contrary to ABNM, the main concept of ABENM is to simulate only infected persons and their immediate contacts at the individual level as agents of the simulation, and to model all other susceptible persons using a compartmental modeling structure. As new persons become infected, their immediate contacts are added as agents (transitioning from the compartmental portion of the model to the network portion of the model), thus evolving the contact network. The key challenge is determining 'who', i.e., the degree (number of contacts), risk group, age, and geographical location of the person, to be added as the immediate contacts of the newly infected node. These characteristics of the infected persons and their contacts are known to be correlated, i.e., persons are more likely to have sexual partners who are similar to them, say of similar age or have similar degree (number of partners) [39]. Correlation in number of partners is a key mathematical feature known as degree correlation between neighboring nodes in scale-free network structure [42], where degree is the number of links (contacts) of a node (person) in the network [31]. In previous work, we developed the ECNA, which uses a neural network model to predict degree correlations, i.e., determine, based on the degree of the newly infected person, the degree of each of their immediate contacts [31]. The ECNA can be used as a network generation algorithm for generation of scale-free networks in ABENM. However, our previous work only modeled hypothetical networks and diseases using simplified data assumptions.

We developed PATH 4.0 model using ABENM and ECNA. Specifically, expanding on the concepts of ECNA, we developed a new network generation model (HIV-ECNA) for simulating HIV

transmissions. For simulating the progression of infection along the HIV disease and care continuum stages for HIV-infected persons, we adopted the disease and care continuum progression model from PATH 2.0 [6]. We implemented PATH 4.0 for prediction of sexually transmitted cases of HIV in the United States over the period 2006 to 2017, among heterosexual female (HET female), heterosexual male (HET male), and men who have sex with men (MSM). We did not model HIV transmissions through injection drug use. We validated the model by comparing epidemic predictions from PATH 4.0 with data from the U.S. NHSS for years 2010 to 2017.

In this paper, we present the development and implementation of PATH 4.0 to the United States. To demonstrate the ability of PATH 4.0 to generate clusters similar to those detected in molecular data, we compare clusters extracted from the PATH 4.0 transmission network to those identified by nucleotide sequencing of HIV genetic data collected from HIV-infected persons by the NHSS [19]. We used a previously developed cluster generation algorithm for extraction of clusters from PATH 4.0 [43].

2. Methods: PATH 4.0 model

In this section, we discuss the structure and methods of PATH 4.0 and its implementation for simulating the HIV epidemic in the United States for the period 2006 to 2017. This section is structured as follows: in Section 2.1, we discuss the overall computational structure of PATH 4.0; in Section 2.2, we discuss the four main modules of PATH 4.0, namely, a compartmental module for simulating susceptible persons, a Bernoulli transmission module for simulating new infections, the HIV-ECNA network generation module for generating sexual partnership networks of newly infected persons, and a disease progression module for simulating HIV-related events for HIV-infected persons; and in Section 2.3, we discuss the implementation of PATH 4.0 for simulating HIV in the United States for the period 2006 to 2017, and provide an overview of the full simulation model. While we present only a limited version of the mathematical concepts and data assumptions and sources here, further details for each section can be found in the corresponding sections of the Appendix. All mathematical notations used in the section are summarized in Table 1. PATH 4.0 was computationally coded in the Netlogo 6.1.0 [44] software, an open-source programmable modeling environment for agent-based modeling with network features.

Table 1. Table of Notations.

Notation	Description
t	Simulation time-step.
A	The number of age-groups.
R	The number of risk-groups.
D	The number of degree bins.
$\mathcal{G}$	The number of pseudo-geographic jurisdictions.
$\bar{a}; \curlyvee; \bar{d}; g$	Used when referring to an age-group, risk-group, degree-bin, and pseudo-geographic jurisdiction, respectively.
$S_t[\bar{a}, \curlyvee, \bar{d}, g]$	An array of size $A \times R \times D \times \mathcal{G}$ representing the number of susceptible persons in the model, in age-group $\bar{a}$, risk group $\curlyvee$, degree-bin $\bar{d}$, and pseudo-geographic jurisdiction g, at time t.
$\mathcal{N}$	A set of nodes, each representing an infected person or a susceptible sexual partner.

Continued on next page

Notation	Description
$\mathcal{E}$	A set of edges representing sexual partnerships between nodes.
$G_t(\mathcal{N}, \mathcal{E})$	A dynamic graph with $\mathcal{N}$ a set of nodes and $\mathcal{E}$ a set of edges, at time t.
Q_t	The number of nodes in graph *G*, at time *t*.
$C_t[i,j]$	A static adjacency matrix of size $Q_t \times Q_t$, with static element $C_t[i,j] = 1$ if i and j are sexual partners anytime during their lifetime and $C_t[i,j] = 0$ otherwise.
$V_t[i,j]$	A dynamic adjacency matrix of size $Q_t \times Q_t$, with element $V_t[i,j] = 1$ if i and j are sexual partners during month t and $V_t[i,j] = 0$ otherwise.
$e = \{i,j\}$	An edge in graph G_t representing a sexual partnership between $nodes\ i$ and j
$\bar{t}(\{i,j\})$	The partnership initiation time; represents the simulation month for partnership initiation.
$\underline{t}(\{i,j\})$	The partnership termination time; represents the simulation month for partnership termination.
$\{\bar{a}_i, \bar{a}_j\}$	The age of nodes i and j at the time of their partnership initiation.
$\{\underline{a}_i, \underline{a}_j\}$	The age of nodes i and j at the time of their partnership termination.
$\bar{a}_{t,j}$	Age-group of node j at time t.
$a_{t,j}$	Age of node *j* at time *t*.
$\bar{d}_j$	Degree-bin corresponding to the number of lifetime partners of node j.
d_j	The actual number of lifetime sexual partners of node j.
$\hat{d}_{\mathrm{t,j}}$	The number of lifetime sexual partners of person j who are already added as nodes in graph *G* at time *t*. For infected nodes $d_j = \hat{d}_{\mathrm{t,j}}$ for susceptible nodes in *G*, $d_j \geq \hat{d}_{\mathrm{t,j}}$
$L_{t,j}$	A partnership distribution matrix of size $A \times 2$, where $L_{t,j}[\bar{a}, 1]$ is the number of partnerships that initiate at age-group $\bar{a}$, and $L_{t,j}[\bar{a}, 2]$ is the number of partnerships that are yet to be assigned. For infected nodes, $L_{t,j}[\bar{a}, 2] = 0, \forall \bar{a}$
$\hbar_{\mathrm{t,j}}$	Infection status of node *j* at time *t*.
$m_{t,j}$	Deceased status of node *j* at time *t*.
r_j	Risk-group of person *j*.
$s_{t,j}$	Care continuum or disease stage of person *j* at time *t*.
$p_{t,j}$	Infectiousness or risk of transmission per act for person *j* at time *t*.
ε	Condom effectiveness.
$s_{t,j}$	The number of sex acts per month for person *j* at time *t*.
$c_{t,j}$	The proportion of acts condom protected of person *j* at time *t*.
$F^{-1}(u)$	The inverse Bernoulli distribution that takes values 1 with probability u and 0 with probability $1 - u$.
D_k	Random variable for degree of node k.
$\Pr(D_k = d_k \mid D_l = d_l)$	Conditional probability distribution for D_k.
$\Pr(D_k = d_k)$	Marginal probability distribution for D_k.
m	Minimum degree of the network.
λ_{r_l}	Scale-free network parameter corresponding to the risk-group of node *l*.
$\bar{L}[\bar{a}, \bar{d}]$	A matrix of size $A \times D$, representing the proportion of partnerships that initiate at age-group $\bar{a}$ for persons in degree-bin $\bar{d}$.

2.1. General structure of PATH 4.0

In this section we present the overall computational structure of PATH to help describe how the ABENM simulation technique combines the features of ABNM and compartmental simulation techniques, but we first briefly describe the main features of this computational structure. Only HIV-infected persons and their immediate contacts (both susceptible and infected contacts) are modeled at the individual-level as agents or nodes in a network (as in an ABNM), and all other susceptible persons are modeled at the population-level (as in a compartmental model). Therefore, at any time point in the simulation, all infected persons are nodes in the network, and all contacts an infected person would have over their lifetime (the contacts may be infected or susceptible) are also nodes in the network. Over time, as new persons become infected, they are added to the network, but at any "current" time point of the simulation, only persons who are currently infected and their contacts (all partners the infected person would have over their lifetime) are nodes in the network. An infected person is connected to each of their lifetime partners through an edge (link), and thus an edge (or link) represents a partnership. Thus, if a node is infected, the model is set up to ensure that their *current degree* (defined as the number of persons they are linked to in the network at that time point in the simulation) is equal to their *actual degree* (defined as the number of partners the infected person will have over their lifetime). However, the links are set up such that each partnership (link) is activated and deactivated over time based on when the partnership initiated and dissolved, through the use of edge features (similar to assigning features such as age to a node, we can assign features to an edge) to keep track of partnership initiation and termination times. As the only susceptible persons who are tracked as nodes in the network are those who are contacts of an infected person, the *current degree* of a susceptible node is less than or equal to their *actual degree* as they are only connected to their infected contacts. Note that it is possible that the susceptible contacts of a newly infected node are already in the network as contacts of other infected persons, but links between two partners are generated only when at least one of them become infected (the methodological process to achieve this is the core of the newly developed HIV-ECNA network generation model and is discussed in Section 2.2.2—in this section we only describe the general computational structure of PATH). All susceptible persons who are not contacts of a currently infected person are in the compartmental model. We will discuss in Section 2.2. the process of determining if a susceptible person in the compartmental model would become a contact of an infected node in the network, and the modeling of their transitioning from the compartmental model to the network.

We next mathematically present the computational structure of PATH 4.0.

Following the compartmental modeling structure, we use a four-dimensional array to keep track of the number of susceptible persons (who are not contacts of a currently infected person), specifically, let,

> S_t = an array of size $A \times R \times D \times \mathcal{G}$, where A is the number of age-groups, R is the number of risk-groups, D is the number of degree-bins (degree is the number of contacts per person and degrees are grouped into bins analogous to age grouped into age-groups), and $\mathcal{G}$ is the number of pseudo-geographic jurisdictions, to model heterogeneity in contact mixing as persons are more likely to form partnerships with persons in the same geographic area (here we only model 'pseudo'-geographic jurisdictions, i.e., we assigned persons to a randomly chosen jurisdiction for purposes of introducing heterogeneity in contact selection, to more realistically represent

network structure formations, but we did not explicitly model geographic features of the epidemic as the focus of this paper is on national aggregated estimates) [45] then

$S_t[\bar{a}, \bar{r}, \bar{d}, g]$ is the number of susceptible persons in age-group $\bar{a}$, risk group $\bar{r}$, degree-bin $\bar{d}$, and pseudo-geographic jurisdiction g, at time t.

Following the agent-based network modeling structure, we use a dynamic graph $G_t(\mathcal{N}, \mathcal{E})$ to track HIV-infected persons and their immediate contacts (these contacts may be infected or susceptible), where $\mathcal{N}$ is a set of nodes, each node representing an infected person or a susceptible sexual partner, and $\mathcal{E}(G_t)$ is a set of undirected edges, an edge $\{i, j\}$ representing a sexual partnership between nodes i and j. The number of nodes in the graph $Q_t = |\mathcal{N}(G_t)|$ and the number of edges $|\mathcal{E}(G_t)|$ are dynamically changing over time t.

The graph $G_t(\mathcal{N}, \mathcal{E})$ has the following features:

Static adjacency matrix: C_t of time-variant size $Q_t \times Q_t$, with static elements $C_t[i, j] = 1$ if i and j are sexual partners anytime during their lifetime and $C_t[i, j] = 0$ otherwise, and

Dynamic adjacency matrix: V_t of time-variant size $Q_t \times Q_t$, with element $V_t[i, j] = 1$ if i and j are in a partnership during month t and $V_t[i, j] = 0$ otherwise.

Each edge $\{i, j\} \epsilon \mathcal{E}$ has the following features (similar to nodes having features of say age, sex, etc., edges can also have features):

Partnership initiation time: $\overline{t}(\{i, j\})$ representing the simulation month for when the partnership initiated,

Partnership termination time: $\underline{t}(\{i, j\})$ representing the simulation month when the partnership terminated,

Partnership initiation age: $\{\bar{a}_i, \bar{a}_j\}$ representing the age of nodes i and j, at the time of partnership initiation, and

Partnership termination age: $\{\underline{a}_i, \underline{a}_j\}$ representing the age of nodes i and j, at the time of partnership termination.

Each node $j \epsilon \mathcal{N}(G_t)$ has the following features:

Actual degree: d_j representing the actual number of lifetime sexual partners of node j,

Current degree: $\hat{d}_{\mathrm{t,j}}$ representing the number of lifetime sexual partners of person j who are already added as nodes in $G_t(\mathcal{N}, \mathcal{E})$; if node j is infected $\hat{d}_{\mathrm{t,j}} = d_j$, if node j is susceptible $\hat{d}_{\mathrm{t,j}} \leq d_j$, and thus dynamically changing with time t,

Partnership distribution matrix: $L_{t,j}$ of size $A \times 2$, where A is the number of age-groups, $L_{t,j}[\overline{a}, 1]$ is the number of partnerships that node j initiates in age group $\overline{a}$, and $L_{t,j}[\overline{a}, 2]$ is the number of partnerships that are yet to be assigned; the sub-script t are to indicate that the values of column 2 of $L_{t,j}$ can change over time, specifically, $L_{t,j}[,2]$ is a column of zeros if the node is infected as all their partnerships are already assigned, and greater than or equal to zero if the node is susceptible (when the susceptible person is added as a contact of a different infected person one of the rows is decremented, and when the susceptible person become infected all rows of column 2 are decremented to zero as their partners are found and added - the HIV-ECNA

was specifically developed for determining when and how to assign these partnerships, and thus generating the network, which is discussed in Section 2.2.3.),
Infection status: $\hbar_{\mathrm{t,j}} = 1$ if node j is an HIV-infected node and $\hbar_{\mathrm{t,j}} = 0$ otherwise,
Deceased status: $m_{t,j} = 1$ if node j is alive and 0 otherwise,
Age: $a_{t,j}$ taking an integer value representative of the age of node j,
Pseudo-geographic jurisdiction: g_j taking an integer value representative of the pseudo-geographic location of node j,
Risk group: r_j taking one of the following values, representative of risk-group of node j, $r_j \in$ {heterosexual female, heterosexual male, MSM}, and
Care continuum and disease stage: $s_{t,j}$ taking one of the following values, 0 (not infected), 1 (infected, acute HIV stage, and undiagnosed), 2 (non-acute HIV, and undiagnosed), 3 (diagnosed and not in care), 4 (in care not on antiretroviral therapy (ART) treatment), 5 (on ART no viral load suppression (VLS)), or 6 (on ART with VLS).

The main relationships between different components of the graph $G_t(\mathcal{N}, \mathcal{E})$ are the following.

Between partnership initiation $\overline{t}(\{i,j\})$ and termination $\underline{t}(\{i,j\})$ times and static and dynamic adjacency matrices (C_t and V_t):
$C_t[i,j] = \begin{cases} 1 \; if \; \{i,j\} \epsilon \mathcal{E} \\ 0 \; otherwise \end{cases}$, i.e., if $\{i,j\}$ are partners at some point during their life, this will have a value of 1,
$V_t[i,j] = \begin{cases} 1 \text{ if } \overline{t}(\{i,j\}) \leq t \leq \underline{t}(\{i,j\}) \\ 0 \; otherwise \end{cases}$, i.e., if $\{i,j\}$ are partners at time t this will have a value of 1, and thus,
$C_t[i,j] \geq V_t[i,j]$.

Between actual degree d_j, current degrees $\hat{d}_{\mathrm{t,j}}$, and static adjacency matrix C_t:

$\hat{d}_{\mathrm{t,j}} \begin{cases} = d_j \text{ if node j is infected} \\ \leq d_j \text{ if node j is susceptible} \end{cases}$, i.e., if a node is infected, they are linked to all partners they will have (actual degree) over their lifetime and thus $d_j = \hat{d}_{\mathrm{t,j}}$, and if a node is susceptible, they are only linked to their infected partners and thus $d_j \leq \hat{d}_{\mathrm{t,j}}$, and

$\hat{d}_{\mathrm{t,j}} = \sum_{i=1:Q_t} C_t[i,j]$, i.e., C_t keeps track of their current degree.

Between actual degree d_j and partnership distribution matrix $L_{t,j}$:
$\sum_{\overline{a}=1:A} L_{t,j}[\overline{a}, 1] = d_j$, at any t, i.e., as $L_{t,j}[\overline{a}, 1]$ tracks number of partnerships that initiate at age-group $\overline{a}$, when summed over all $\overline{a}$ it should add to the actual degree d_j for all nodes whether infected or susceptible, and

$\sum_{\overline{a}=1:A} L_{t,j}[\overline{a}, 2] = \begin{cases} d_j - \hat{d}_{\mathrm{t,j}} \text{ if node } j \text{ is susceptible} \\ 0 \text{ if node } j \text{ is infected} \end{cases}$, i.e., as $L_{t,j}[\overline{a}, 2]$ tracks number of partnerships that initiate at age-group $\overline{a}$ and are yet to be generated, $\sum_{\overline{a}=1:A} L_{t,j}[\overline{a}, 2]$would be zero if the node is infected because all partnerships of an infected node are already connected

in the network, and would be equal to the number of partners yet to be assigned if the node is susceptible. (Assigning partnerships and all other features related to the network are part of the newly developed HIV-ECNA network generation algorithm, discussed later).

This section presented the computational structure of the model, specifically the compartmental modeling structure, the network structure, and the features of the nodes and edges in the network. A visual representation of the computational structure is presented in Figure 1(A). The next section describes the methods (modules) used in modeling these features, and Figure 1(B) provides an overview of the modules.

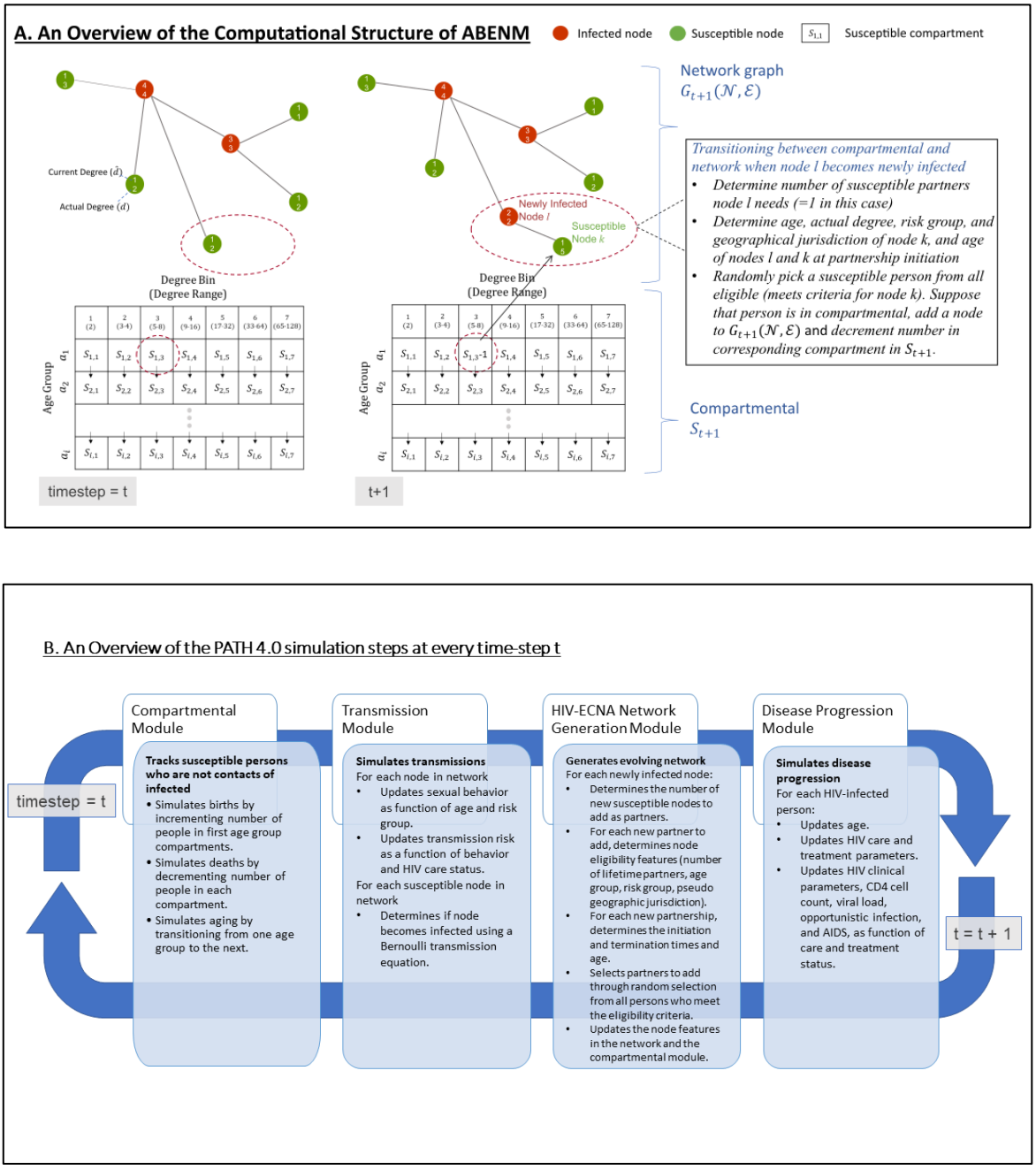

Figure 1. Schematic overview of the computational structure of ABENM (A) and simulation steps (B).

2.2. Four main modules of PATH 4.0

At every time-step (monthly) of the simulation, the model runs four different modules: a compartmental module for simulating susceptible persons, a Bernoulli transmission module for simulating new infections, the HIV-ECNA network generation module for generating sexual partnership networks of newly infected persons, and a disease progression module for simulating HIV-related events for HIV-infected persons. We discuss each module below.

2.2.1. Compartmental module for simulating susceptible persons

Every time-step (monthly) of the simulation, this module updates the demographic features of susceptible persons tracked through the array S_t. Specifically, it simulates births, aging, and deaths as follows

$$S_t[\bar{a}, r, \bar{d}, g] = S_{t-1}[\bar{a}, r, \bar{d}, g] + \frac{dS[\bar{a}, r, \bar{d}, g]}{dt}\Delta t$$

$$\frac{dS[\bar{a}, r, \bar{d}, g]}{dt} = \begin{cases} B_{r,\bar{d}} - \mu_{\bar{a}} S_{t-1}[\bar{a}, r, \bar{d}, g]; & if\ \bar{a} = the\ youngest\ age\ group\ 13-17 \\ -\left(\mu_{\bar{a}} + \frac{1}{|\bar{a}|}\right) S_{t-1}[\bar{a}, r, \bar{d}, g] + \left(\frac{1}{|\bar{a}-1|}\right) S_{t-1}[\bar{a}-1, r, \bar{d}, g]; & otherwise \end{cases}$$

where,

$B_{r,\bar{d}}$ is the number of persons of risk group r and degree-bin $\bar{d}$ annually aging into the youngest age-group $\bar{a}$ =13-17 years, we assumed equal birth rate for all pseudo-geographic areas,

$\mu_{\bar{a}}$ is the annual natural mortality in age-group $\bar{a}$.

$|\bar{a}|$ is the age-group interval size of age-group $\bar{a}$, and thus $\frac{1}{|\bar{a}|}$ is the rate of aging out.

Δt is 1/12 to represent the modeling of monthly time-steps.

In the simulation, we estimate $B_{r,\bar{d}}$ as the number of persons who age into age-group 13-17 years multiplied by the proportion of persons of risk-group r ($r \in$ {heterosexual male, heterosexual female, MSM}), and further multiplied by the proportion of persons with degree-bin $\bar{d}$ (where $\bar{d} \in \{1,2, \ldots D\}$). Values for the proportions in risk-group and degree-bin are specific to the population simulated and are discussed in the Appendix for application to the US population. We use log-2 binning for degree, i.e., persons in degree-bin $\bar{d}$ are those with lifetime number of partners (d) in $2^{\bar{d}-1} < d \leq 2^{\bar{d}}$. As the number of lifetime sexual partners follows a power-law distribution [34], i.e., $Pr(d = k) \sim k^{-\lambda}$, where λ is the scale-free parameter of the distribution, following the characteristic feature of power-law distributions, it would mean that a large number of persons have lower degrees and only a few persons have a very high degree. This creates issues when using uniform binning. For example, persons with large number of partners might typically report rounded numbers, e.g., 50, 100, instead of 48 or 98 partners, so using uniform bins of say width 5 or 10 would create spikes at rounded values and zero around it. Therefore, as commonly done, for degree-bins we use log-2 binning, i.e., persons in degree-bin $\bar{d}$ are those with lifetime number of partners (d) in $2^{\bar{d}-1} < d \leq 2^{\bar{d}}, \bar{d} \in \{1,2, \ldots D\}$, which would create bins of narrower intervals for smaller degree and wider intervals for larger degree. Applying the power-law distribution, we calculate the proportion of persons in degree-

bin $\bar{d}$ as $\left[\sum_{k=2^{\bar{d}-1}+1}^{2^{\bar{d}}} k^{-\lambda}\left(\sum_{k=2^0+1}^{2^D} k^{-\lambda}\right)^{-1}\right]$, using the value of λ specific to the population simulated as discussed in the Appendix.

2.2.2. Transmission module for simulating new infections

Every time-step (monthly) of the simulation, this module determines if a susceptible node l in graph $G_t(\mathcal{N},\mathcal{E})$ becomes infected using a Bernoulli transmission equation. Note, as susceptible persons in the compartmental model array S_t are not connected to an infected person, their chance of infection is zero. Further note that, susceptible persons can move from compartmental model to the network upon becoming partners of the infected person, modeled using the HIV-ECNA algorithm discussed in the next section, which would then expose them to the infection.

Specifically, for nodes in the network with HIV infection status $\hbar_{\mathrm{t}-1,l} = 0$ (denoting susceptible) the Bernoulli transmission equation is used to estimate the updated value $\hbar_{\mathrm{t},l}$ as follows.

$\hbar_{\mathrm{t},l} = F^{-1}\left(1 - \prod_{j=1}^{Q_t}\left(1-\alpha_j\varepsilon\right)^{s_{t,j}\cdot c_{t,j}}\left(1-\alpha_j\right)^{s_{t,j}\cdot(1-c_{t,j})}\right)$, where,

$\alpha_j = V_t[l,j].\hbar_{\mathrm{t,j}}.m_{t,j}.p_{t,j}$, where $V_t[l,j], \hbar_{\mathrm{t,j}}, m_{t,j}$ are the elements of the graph described in Section 2.1, and $p_{t,j}$ is the probability of transmission per act modeled as a function of disease and care stage $s_{t,j}$ and risk group r_j of the infected node j; we will have a value of $\alpha_j = p_{t,j}$ if j is a contact of l (i.e., $V_t[l,j] = 1$), is infected (i.e., $\hbar_{\mathrm{t,j}} = 1$), and is alive (i.e., $m_{t,j} = 1$), and $\alpha_j = 0$ otherwise,

ε = condom effectiveness,

$s_{t,j}$ = number of sex acts per month with node j, modeled as a function of age, risk group, and number of partners of node j,

$c_{t,j}$ = proportion of acts with node j that is condom protected, modeled as a function of age, risk group, and number of partners of node j,

$F^{-1}(u) =$ an inverse Bernoulli distribution that takes a value of 1 with probability u and value of 0 with probability $1-u$.

If node l becomes infected, i.e., if the above equations yield a value of 1 ($\hbar_{\mathrm{t},l} = 1$), we set its HIV stage as 1 (i.e., set $s_{t,l} = 1$ to denote the first stage of HIV, which is acute and undiagnosed).

Every time-step t, this module also determines and updates any changes in sexual behavior of infected nodes. Specifically, it updates the following values. For every partnership (k,j), its active/inactive status by checking if the current time-step is within the partnership initiation and termination time, as $V_t[k,j] = \begin{cases} 1 \text{ if } \bar{t}(\{k,j\}) \le t \le \underline{t}(\{k,j\}),\ indicating\ it\ is\ active \\ 0\ otherwise, indicating\ it\ is\ inactive \end{cases}$. For every infected node j, the number of sex acts per month as a random draw from age-group and risk-group specific uniform distribution, corresponding to age $a_{t,j}$ and risk group r_j of node j. For every infected node j, the number of sex acts per partner $s_{t,j}$ as $s_{t,j} = \frac{\text{number of sex acts per month for node } j}{\sum_{k=1:Q_t} V_t[k,j]}$. For every infected node j, condom use $c_{t,j}$ as a function of number of active partners ($\sum_{k=1:Q_t} V_t[k,j]$) and disease stage $s_{t,j}$ (specifically, diagnosed status of HIV, with higher condom use if aware, i.e., $s_{t,j} > 2$. For every infected node j, the infectiousness $p_{t,j}$ as a function of risk group r_j and stage $s_{t,j}$ of node j. Data

assumptions for the above behavioral parameters are presented in Section S3 of the Appendix, specific to the US.

2.2.3. HIV-ECNA network generation module for generating partnerships of newly infected nodes

This module controls the overall network dynamics of partnerships between nodes. It controls partnership formation and dissolution over time and age, modeled through a combination of the static (C_t) and dynamic (V_t) adjacency matrices, with C_t keeping track of all partnerships over the lifetime and V_t keeping track of only those that are active at that specific time. It controls the dynamics between age-group and risk-group mixing between partnerships. It also controls the transitioning of susceptible persons from the compartmental model S_t to the network $G_t(\mathcal{N}, \mathcal{E})$ as they become partners of newly infected persons. These network dynamics are modeled through the simulation of HIV-ECNA, which we discuss next.

At the beginning of every time-step $\mathrm{t}+1$, this module applies the HIV-ECNA for each node l that was newly infected at the end of the previous time-step t. As l was a susceptible person up until time-step t, it had links with only their infected partners, and thus their current degree ($\hat{d}_{\mathrm{t,l}}$) was less than or equal to their actual degree (d_l). Therefore, the main functionality of this algorithm is to generate the contact network for each newly infected node l, i.e., determine that $d_l - \hat{d}_{\mathrm{t,l}}$ partners are yet to be assigned, determine who those persons are (including the degree-bin corresponding to their number of lifetime partners, current age-group, risk-group, and pseudo-geographic jurisdiction) and, if they are not already part of the graph $G_t(\mathcal{N}, \mathcal{E})$, add them to $G_{t+1}(\mathcal{N}, \mathcal{E})$ and remove them from S_{t+1}.
The steps of the HIV-ECNA are as follows.

For every newly infected node l,

1. Determine the number of new partnerships (edges) to generate as actual degree minus current degree (i.e., $d_l - \hat{d}_{\mathrm{t,l}}$). Note, these new partnerships would all be with susceptible persons.
2. For each new susceptible partner node, say k, determine node features, specifically, its number of lifetime partners degree-bin $\bar{d}_k$, risk-group r_k, and current age-group $\bar{a}_{t,k}$, pseudo-geographic jurisdiction g_k, and the partnership distribution matrix $L_{t,k}$.
3. For each new edge (partnership) between l and k, say $\{l, k\}$, determine its edge features, specifically, the partnership initiation age $\{\bar{a}_l, \bar{a}_k\}$, initiation time $\bar{t}(\{l, k\})$, termination age $\{\underline{a}_l, \underline{a}_k\}$, and termination time $\underline{t}(\{l, k\})$.
4. Determine who each new partner k is by a uniform random draw from all who are eligible, i.e., all persons who are eligible have an equal chance of selection. All susceptible agents in the graph $G_t(\mathcal{N}, \mathcal{E})$ and susceptible non-agents in the compartmental model array S_t with node and edge features matching that in steps 2 and 3 above, can be eligible
   a. A susceptible agent, say j, is eligible if its current age $a_{t,j} \in \bar{a}_{t,k}$, its risk group $r_j = r_k$, its actual degree $d_j \in \bar{d}_k$, its pseudo-geographic jurisdiction $g_j = g_k$, its degree minus

current degree $d_j - \hat{d}_{t,j} > 0$, and the number of unassigned contacts in age-group $\bar{a}$, corresponding to activation age $\bar{a}_k$ is greater than zero, i.e., $L_{t,j}[\bar{a}, 2] > 0, \bar{a}_k \in \bar{a}$.

b. All susceptible non-agents in element $S_t[\bar{a}_k, r_k, \bar{d}_k, g_k]$ are eligible.

Therefore, the probability that the person picked is a susceptible agent is the number of eligible agents divided by the number of eligible agents and non-agents, and thus, as the network grows, the chance of selection from within the network increases.

5. For each new partner k (determined in previous step) and partnership $\{l, k\}$, update their corresponding features in $G_t(\mathcal{N}, \mathcal{E})$ and S_t, i.e., update all elements of the computational structure described under Section 2.1, generating an updated graph $G_{t+1}(\mathcal{N}, \mathcal{E})$ and an updated array S_{t+1}
   a. If the new partner k is already a susceptible node in graph $G_t(\mathcal{N}, \mathcal{E})$ update the graph features corresponding to nodes l and k and add a new edge $\{l, k\}$.
   b. If the new partner k is an element of the compartmental array S_t, add a new node k to the graph, update the graph features corresponding to nodes l and k, add a new edge $\{l, k\}$, and decrement the value of $S_t[\bar{a}_k, r_k, \bar{d}_k, g_k]$ by 1 (thus transitioning the susceptible person from compartmental model to the network). Assign actual degree d_k through random selection from the degree-bin $\bar{d}_k$ and assign current age as $a_{t,k} = \bar{a}_k - (\bar{t}(\{l, k\}) - t)$, where t is the current time-step

Below, we briefly discuss the methods for determining the features in steps 2 and 3 and provide further details in the Appendix Section S4.

Determining degree-bin $\bar{d}_k$ for each new partner k

As described in Step 1, suppose it is determined that a new susceptible person (say k) should be added as the partner of the newly infected node l. A key aspect of the HIV-ECNA is determining the features of k, including its degree-bin (the bin corresponding to its number of lifetime partners) $\bar{d}_k$. For a node l, the degree-bin of a partner $\bar{d}_k$ is not independent of its degree-bin $\bar{d}_l$ because of degree correlations between node neighbors, a key feature of scale-free networks [42]. Generally speaking, this means that the degree-bin $\bar{d}_k$ should be determined using some probability distribution that is dependent on the degree-bin of the infected neighbor $\bar{d}_l$, that is, degree $\bar{d}_k$ is conditional on $\bar{d}_l$. Mathematically representing this, $\Pr(\bar{D}_k = \bar{d}_k | \bar{D}_l = \bar{d}_l) \neq \Pr(\bar{D}_k = \bar{d}_k)$, and thus, $\bar{d}_k$ cannot be directly drawn from its scale-free network probability mass function $\Pr(\bar{D}_k = \bar{d}_k) \sim \sum_{i=2^{\bar{d}_{k-1}+1}}^{2^{\bar{d}_k}} i^{-\lambda}$ but should be determined using a conditional probability distribution $\Pr(\bar{D}_k = \bar{d}_k | \bar{D}_l = \bar{d}_l)$; where $\bar{D}_k$ is the random variable for degree-bin of node k.

While the literature presents an analytical method for estimation of $\Pr(\bar{D}_k = \bar{d}_k | \bar{D}_l = \bar{d}_l)$ for general static scale-free networks [42], our previous work showed that this method is not suitable in the context of simulating an epidemic in a dynamically evolving contagion network [31]. Specifically, in general static scale-free networks, the full network is available so the degree of all node neighbors are available, and thus $\Pr(\bar{D}_k = \bar{d}_k | \bar{D}_l = \bar{d}_l)$ is an expectation over all possible values of $\bar{d}_l$ i.e., an average over "all" node neighbors. However, as we only simulate infected nodes and their immediate

contacts in the network, in our context, only values corresponding to infected node neighbors are used. As it is more likely that nodes with higher degree get infected first, the value of $\Pr(\bar{D}_k = \bar{d}_k | \bar{D}_l = \bar{d}_l)$ when $l = all\ node\ neighbors$ is different compared to when $l = infected\ node\ neighbors$ [34]. And further, $\Pr(\bar{D}_k = \bar{d}_k | \bar{D}_l = \bar{d}_l)$ is likely to change over time as the epidemic spreads and the percent of the population that is infected changes. Our previous work developed a neural network model for the prediction of $\Pr(\bar{D}_k = \bar{d}_k | \bar{D}_l = \bar{d}_l)$ using as independent variables, $\bar{d}_l$, $\bar{d}_k$, the minimum degree of the network m, the percent of the population that is infected, and the scale-free network parameter λ_{r_l} corresponding to risk group of node l (r_l). We used a similar method here. Further details of the training of the neural network and data assumptions and sources for the US are presented in the Appendix Section S4.1.

Determining risk group r_k and pseudo-geographic jurisdiction g_k for each new partner k

For any node l, the risk group of partner k (r_k) was determined based on a risk group mixing probability matrix, which is a square matrix of probabilities, with rows and columns representing heterosexual male, heterosexual female, and MSM, and each element, say (r_l, r_k), representing the probability a person in risk group r_l partners with a person in risk group r_k. The pseudo-geographic jurisdiction (g_k) was determined based on a pseudo-geographic mixing probability matrix. Data assumptions and sources for risk group mixing and pseudo-geographic mixing are presented in the Appendix Section S6.5

Determining the partnership distribution matrix $L_{t,k}$ for each new partner k

Suppose d_k is the actual degree (number of lifetime partners) of a newly added susceptible node k. We need to determine at what age of k will each partnership initiate. As these data are not directly available, we estimated it using Markov process-based simulation and optimization methods applied to data from behavioral surveys that reported the number of lifetime partners by persons' age at the time of survey. More specifically, we can describe the parameter of interest and its estimation process as follows. Suppose there is a matrix $\bar{L}$ of size $A \times D$, with A the number of age-groups and D the number of degree-bins, and element $\bar{L}[\bar{a}, \bar{d}]$ representing the proportion of partnerships that initiate at age-group $\bar{a}$ for persons in degree-bin $\bar{d}$, with each column of $\bar{L}$ adding to 1, for all $\bar{d} \in \{1,2, \dots D\}$. Then, for any node k with actual degree $d_k \epsilon \bar{d}$, we can calculate the partnership distribution matrix, i.e., the number of partnerships that initiate at age $\bar{a}$, as $L_{t,k}[\bar{a}, 1] = \bar{L}[\bar{a}, \bar{d}].\, d_k$.

Direct data for $\bar{L}$ would be a longitudinal survey over the duration of life of an individual, where the individual reports the number of partnerships they initiated at every age point of their life. Such surveys, however, are unavailable. Therefore, we estimated $\bar{L}$ using survey data on the reported number of partners up to that time by persons of different age-groups (see Appendix Section 4.2). Note that, these survey data only represent the number of partners up to the current age of the surveyed individual. Thus, the degree-bin $\bar{d}$ each person would belong to is unknown as $\bar{d}$ represents the number of partners the person would eventually have over their full lifetime. The age at which each partnership initiated is also unknown. Therefore, we estimated $\bar{L}[\bar{a}, \bar{d}], \forall \bar{a}, \forall \bar{d}$, by mathematically piecing together

data from persons of different age-groups to mimic a longitudinal survey. We used a two-step process in this estimation of $\bar{L}$. In step 1, we solved for the probabilities of initiating a new partnership when a person ages from age-group $\bar{a}-1$ to $\bar{a}$, by formulating it as the transition probability matrix of a Markov chain and solving for it using an optimization model, calibrating to the survey data. In step 2, we used the transition probability matrix from step 1 to simulate partnership changes over the lifetime of a person, starting from age-group $\bar{a}=1$ to age-group $\bar{a}=A$, repeating for 10,000 people, grouping together all persons who end in the same degree-bin $\bar{d}$ at the last age-group A, and for each degree-bin $\bar{d}$, calculating the average proportion of partnerships that initiated in age-group $\bar{a}$, $\forall\bar{a}$. We discuss both steps in detail in the Appendix Section S4.2.1.

Determining partnership initiation and termination age and time for edges $\{l,k\}$ with each partner k

Determining the expected partnership initiation age $\{\bar{a}_l,\bar{a}_k\}$, initiation time $\bar{t}(\{l,k\})$, termination age $\{\underline{a}_l,\underline{a}_k\}$, and termination time $\underline{t}(\{l,k\})$ of each edge $\{l,k\}$ between the newly infected node l and each of its partner k is equivalent to assigning values of these variables to each of the k partners over the duration of life of the infected node l. The optimal solution are the values that maintain the probability distribution for age-mixing between partners, and maintain the partnership distribution matrices ($L_{t,l}$ and $L_{t,k},\forall k$) of the newly infected node l and each partner k, i.e., for any node i, the number of edges initiating at age-group $\bar{a}$ should be equal to $L_{t,i}[\bar{a},1]$, estimated in the previous section. We formulated this problem as an optimization model and developed a heuristic solution algorithm. The details of the formulation and the heuristic solution algorithm are presented in Appendix Section S4.3. The current age of partner k ($a_{t,k}$) is then determined as $a_{t,k}=\bar{a}_k-(\bar{t}(\{l,k\})-t)$, and $\bar{a}_{t,k}$, such that, $a_{t,k}\in\bar{a}_{t,k}$is set as its current age-group.

2.2.4. Disease progression module for simulating progression along disease and care continuum stages

The disease progression module in PATH 4.0 is similar to that in PATH 2.0 [6]. At every time-unit of the simulation, this module updates the individual-level demographic and disease dynamics for every HIV-infected person, including aging, HIV-related and natural mortality, HIV disease progression, and changes in diagnosis, care, and treatment status.

Updating disease progression includes updating HIV-specific parameters such as CD4 cell count, viral load, opportunistic infection (OI) incidence, and onset of acquired immune deficiency syndrome (AIDS), using previously validated disease progression methods in PATH 2.0. These HIV-specific parameters are updated as a function of care and treatment status and ART regimen. For persons on ART treatment, it also simulates changes in ART regimen over time by simulating viral load rebound. We provide an overview of the disease progression methods and data assumptions and sources in the Appendix Section S5.1–S5.3.

Updating changes in diagnosis, care, and treatment status includes generating events of testing, initiation of treatment, and dropping-out or re-entry into treatment by calibrating to match surveillance estimates for the distribution of PWH by care continuum stages (unaware, aware not in care, and on ART treatment with viral load suppression) corresponding to the population and year being simulated. The details of the calibration method are presented in the Appendix Section S5.4.

2.3. Implementation of PATH 4.0 for simulation of HIV in the US

In application of the proposed ABENM to HIV, we first generated an initial population that is representative of people living with HIV (PWH) in the United States in 2006. We achieved this through a methodology that includes two sequential dry runs of the simulation (the concepts of dry runs for model initialization are discussed in more detail in the Appendix S6.2). The first dry run initializes a network of HIV-infected persons and immediate contacts that is replicative of the contact network among PWH, i.e., matching the correlations in degree, age, and risk group between partners. The second dry run initializes the model with epidemic and demographic features, such as disease stage, care continuum stage, age, and risk group, that are representative of the distributions of these features in PWH in the United States in 2006. To create a representation of HIV in the United States, data were taken from several studies, demographical, sexual behavioral, clinical, and HIV care and treatment behavioral studies, originating from data collected as part of multiple large national surveillance and survey systems in the United States, and other small studies. The surveillance and survey systems include the National HIV Surveillance System (NHSS), the Medical Monitoring Project (MMP), the HIV Outpatient Study (HOPS), the American Community Survey (ACS), the National HIV Behavioral Surveillance (NHBS), the National Survey for Family Growth (NSFG), and the National Survey for Sexual Health and Behavior (NSSHB) [46–52]**.** The specific data sources are detailed in the Appendix, in S2 and S6 for demographics, in S3 and S4 for sexual behavior, and in S5 for clinical and HIV care and treatment behaviors.

After initialization of the model to 2006, we ran the simulation, from 2006 to 2017 in monthly-time steps, by calibrating to the care continuum distribution of PWH in the year being simulated, taking estimates from NHSS. We present the data assumptions and sources in the Appendix Section S6. An overview of the steps of the full simulation model, including the dry runs, is presented in Appendix Section S7.

3. Model validation

3.1. Validation of epidemic predictions

To validate the epidemic predictions from the model, for the period 2010 to 2017, we compared simulated annual estimates of relevant HIV parameters, including total prevalence, diagnosed prevalence, annual incidence, and annual diagnoses, distributed by risk group and age, with surveillance data [53–60]. We define total prevalence as the number of people living with diagnosed or undiagnosed HIV, diagnosed prevalence as the number of people living with diagnosed HIV, annual incidence as the number of new infections in that year, and annual diagnoses as the number of new diagnoses in that year. Risk groups include heterosexual females, heterosexual males, and men who have sex with men (MSM) infected with HIV through sexual transmission. Specifically, we compared model and surveillance estimates for the following features:

1. Distribution of overall disease burden by risk group (Figure 2), calculating for each risk group,
    a. total prevalence in risk group divided by overall total prevalence
    b. diagnosed prevalence in risk group divided by overall diagnosed prevalence
    c. annual incidence in risk group divided by overall annual incidence
    d. annual diagnoses in risk group divided by overall annual diagnoses

2. Measures of epidemic growth within each risk group (Figure 3) as,
    a. annual incidence in risk group divided by total prevalence in risk group
    b. annual diagnoses in risk group divided by diagnosed prevalence in risk group
3. For each risk-group, measures of epidemic growth within each age-group and distribution of disease burden by age (heterosexual females in Figure 4, heterosexual males in Figure 5, and MSM in Figure 6) as,
    a. annual incidence in age-group divided by total prevalence in age-group
    b. annual diagnoses in age-group divided by diagnosed prevalence in age-group
    c. annual incidence in age-group divided by total annual incidence in risk group
    d. annual diagnoses in age-group divided by total annual diagnoses in risk group

We generated 100 runs of the simulation and present box plots marking the minimum, 1st quartile, 2nd quartile, 3rd quartile, and maximum, of the 100 runs for each of the above features along with the corresponding values calculated using surveillance estimates [36].

The surveillance estimates fall within the range of model estimates in most cases. In a few cases, such as in the ratio of new infections to total prevalence for MSM (Figure 3) and distribution of diagnosed cases by age for MSM (Figure 6), the surveillance estimates were outside the range of model estimates in some years. However, we believe the overall results are acceptable considering that these metrics are an outcome of multiple interacting events, including those related to sexual behavior, HIV-related care behavior, and disease progression, each event simulated at the individual-level using age, and risk group specific parameters. Specifically, sexual behavioral events include partnership formation and dissolution over time varying by age and risk group, age-group and risk-group mixing between partnerships, and changes in sexual exposures and condom use varying by age, risk group, and number of partners. HIV-related care events include HIV testing and diagnosis, linkage to care and treatment, and dropping-out of and re-entry into care and treatment. Disease progression events include changes in CD4 cell counts, viral load, OI and AIDS incidence, and mortality, each influenced by time of diagnosis and initiation of treatment, and changes in treatment regimen over time. Therefore, considering that the resulting overall population-level epidemic features presented in Figures 2 through 6 are an outcome of interactions between the above multiple events modeled at the individual-level, and that the model results are close to the surveillance estimates on most of these epidemic features (features not used in model calibration), we believe the model provides an acceptable replication of the epidemic. Further, as the surveillance estimates for incidence and prevalence are nationally aggregated average estimates which also have a range of uncertainty associated with them, we did not want to risk overfitting.

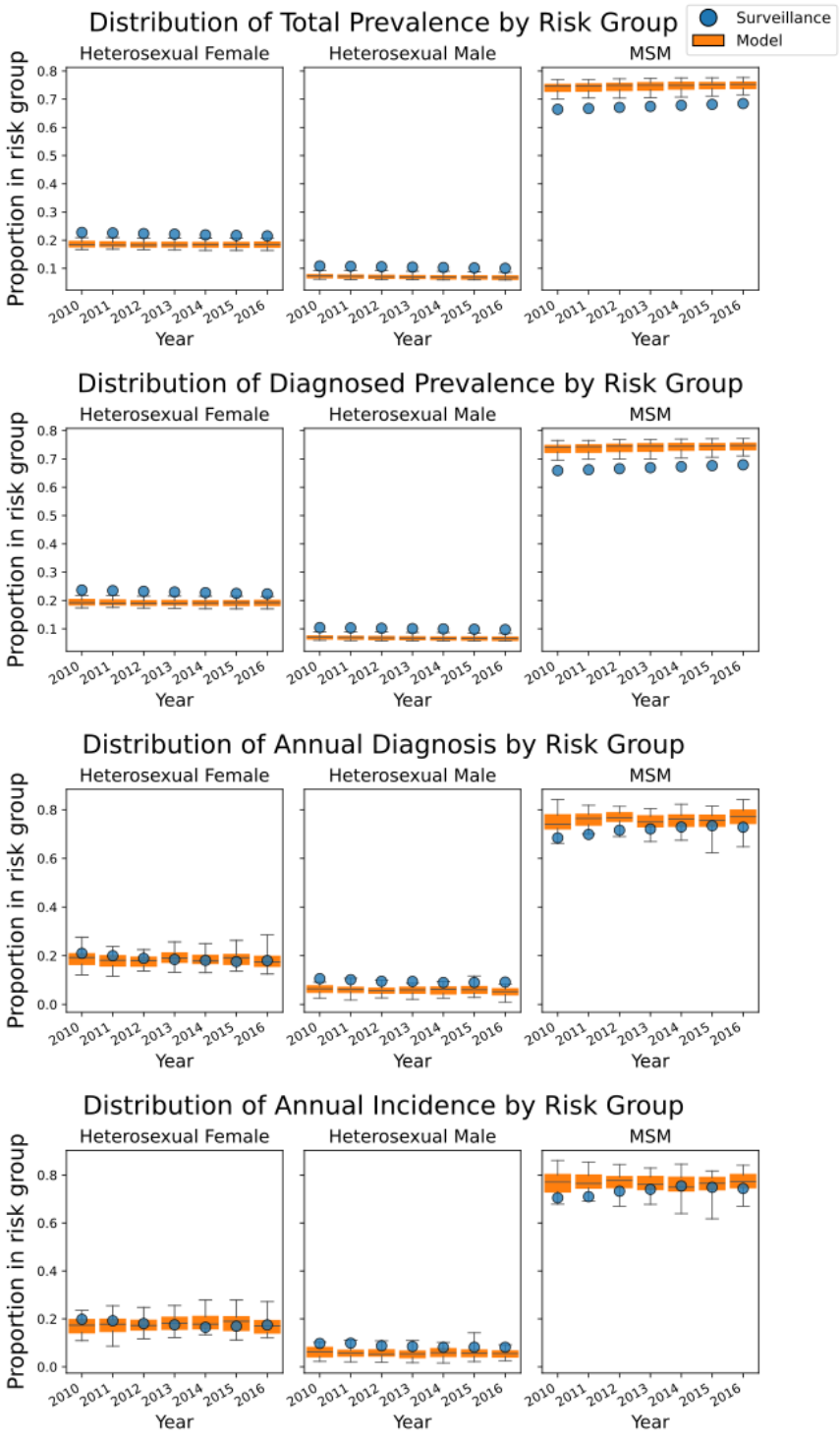

Figure 2. Distributions of total prevalence, diagnosed prevalence, annual incidence, and annual diagnoses by risk group; total prevalence = the number of people living with diagnosed or undiagnosed HIV, diagnosed prevalence = the number of people with diagnosed HIV, annual incidence = the number of new infections in that year, and annual diagnosis = the number of new diagnosis in that year.

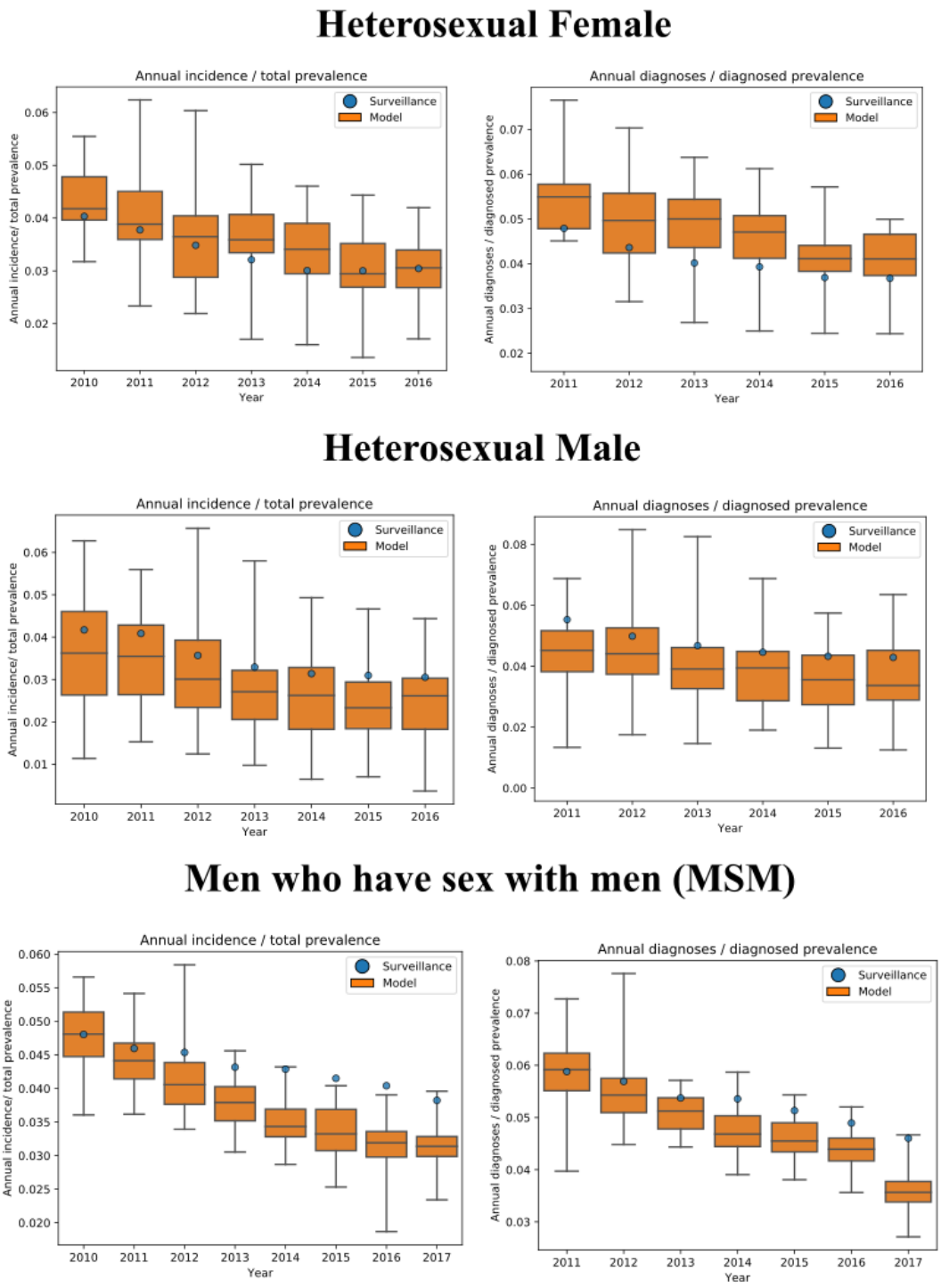

Figure 3. Comparing model estimates with surveillance for annual incidence by total prevalence and annual diagnosis by diagnosed prevalence for each risk group; total prevalence = the number of people living with diagnosed or undiagnosed HIV, diagnosed prevalence = the number of people with diagnosed HIV, annual incidence = the number of new infections in that year, and annual diagnosis = the number of new diagnosis in that year.

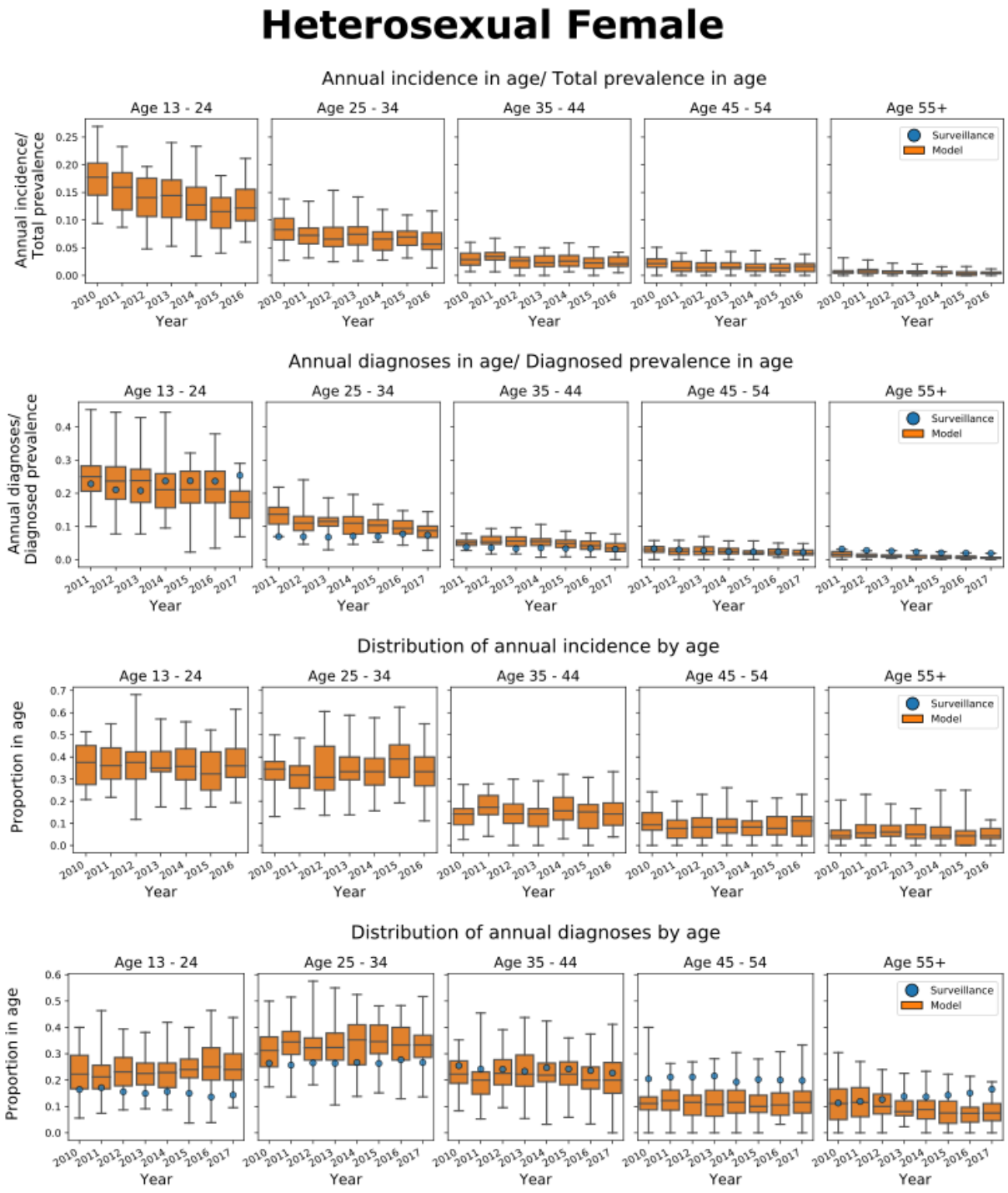

Figure 4. Annual incidence by total prevalence, annual diagnoses by diagnosed prevalence, and distributions of annual incidence and annual diagnoses by age among heterosexual females. Note: charts with no surveillance points are those for which data are available for MSM but not heterosexuals, but we are reporting simulated estimates for completeness.

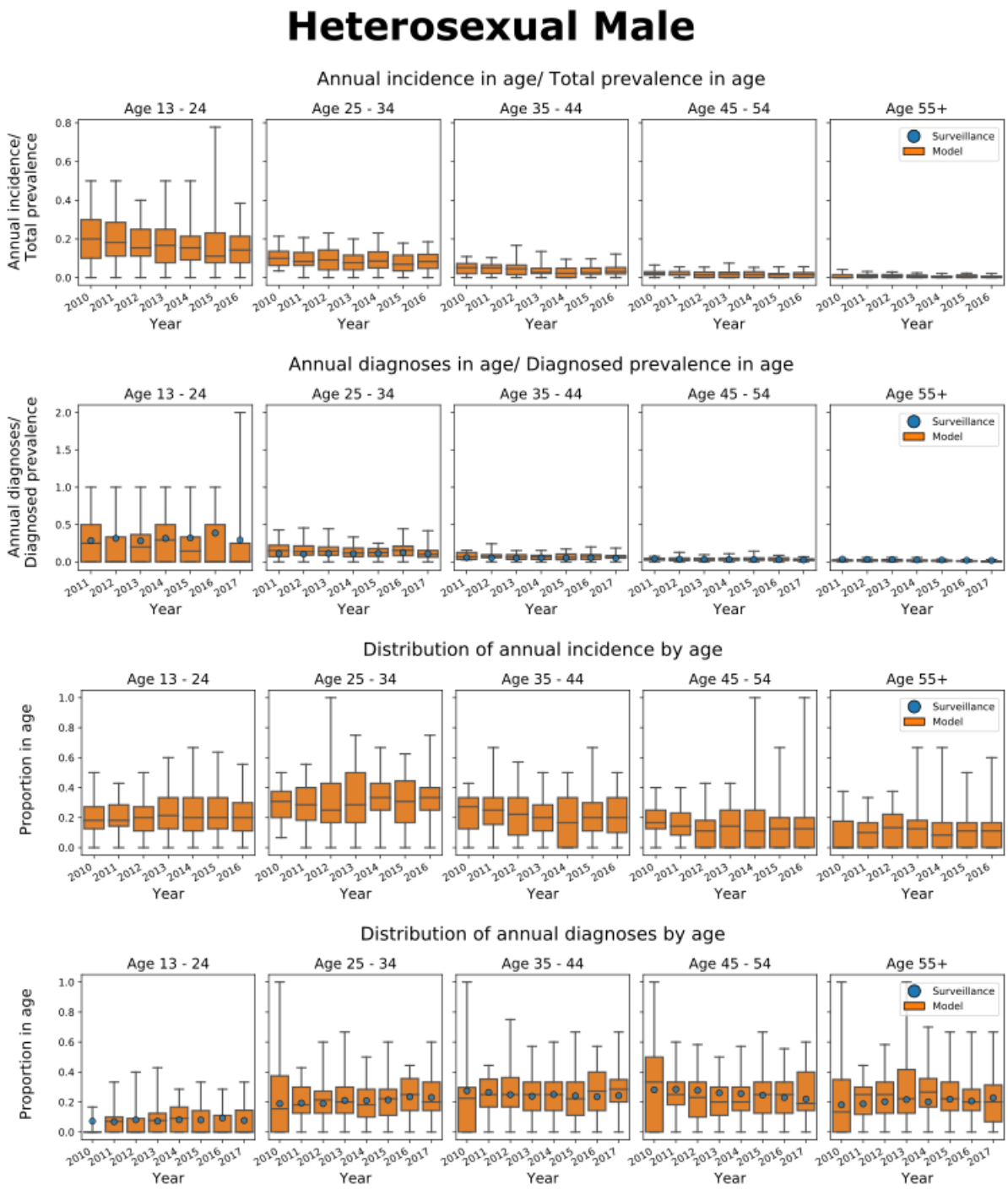

Figure 5. Annual incidence by total prevalence, annual diagnoses by diagnosed prevalence, and distributions of annual incidence and annual diagnosis by age among heterosexual males. Note: charts with no surveillance points are those for which data are available for MSM but not heterosexuals, but we are reporting simulated estimates for completeness.

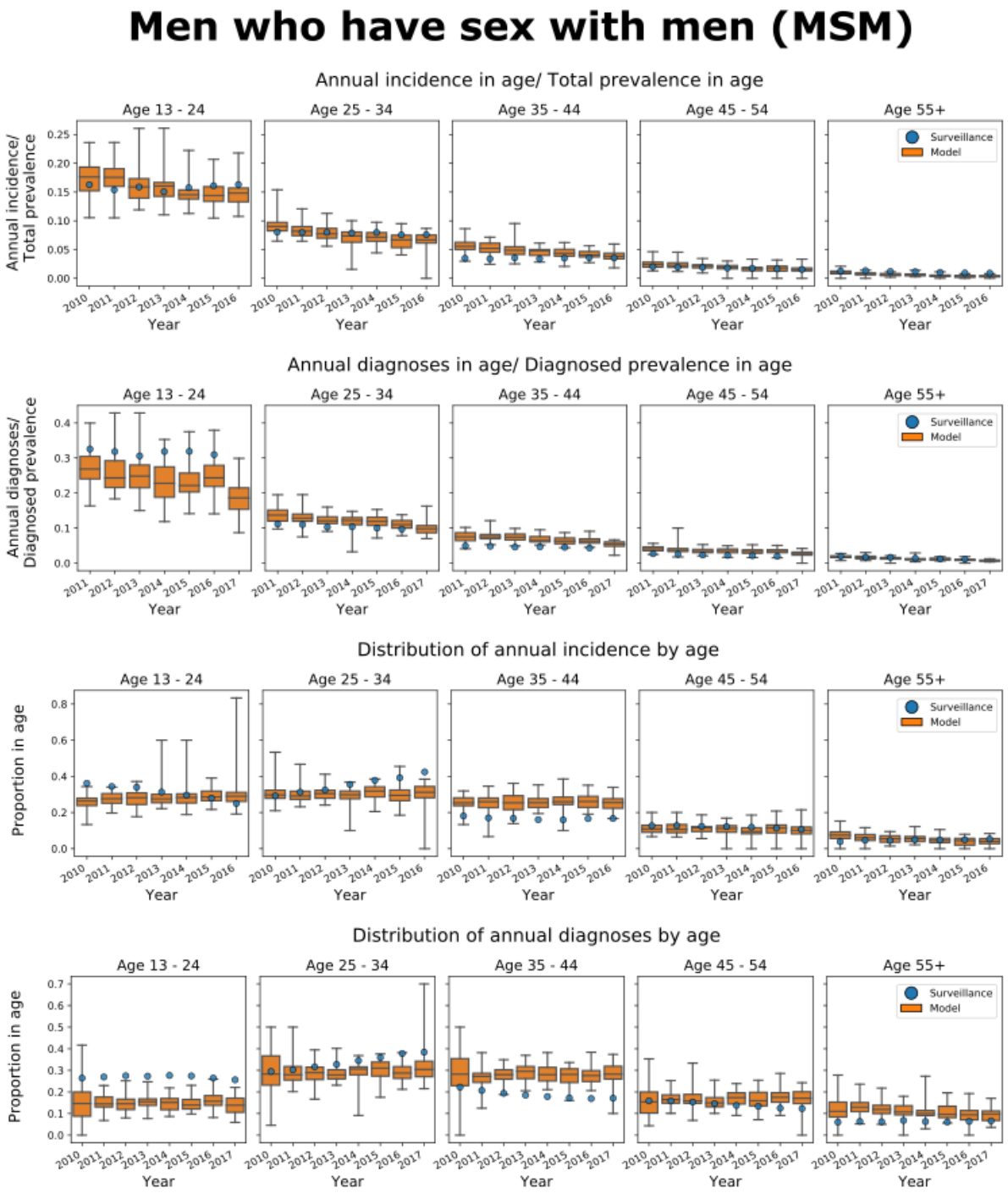

Figure 6. Annual incidence by total prevalence, annual diagnoses by diagnosed prevalence, and distributions of annual incidence and annual diagnosis by age among MSM (men who have sex with men).

4. Testing applicability of path to cluster generation

To test the ability of PATH 4.0 to generate clusters similar to those detected through analysis of nucleotide sequence data reported to NHSS, we compared clusters extracted from the PATH 4.0 transmission network to those identified through molecular cluster analysis of NHSS data. During 2013–2017, 27 HIV surveillance jurisdictions in the United States reported nucleotide sequences from routine clinical drug resistance tests to the NHSS; these jurisdictions reported 70% of US HIV diagnoses in 2015, To identify molecular clusters in NHSS data, we analyzed data reported through December 2017 for HIV infections diagnosed during January 1, 2015–December 31, 2017. Clusters were identified using methods described previously [19]: in brief, we included partial *pol* (protease and reverse transcriptase) sequences that were ≥500 nucleotides in length and removed sequences identified as potential contaminant. Sequences were analyzed using a local installation of HIV-TRACE (HIV TRAnsmission Cluster Engine [61], www.hivtrace.org) following methods previously described [19]. Clusters were defined as connected components using a genetic distance threshold of 0.5%. Sequence data were available for 48% of persons with HIV infection diagnosed during 2015–2017 in the participating jurisdictions.

To replicate molecular cluster detection in the simulation, we applied a cluster generation algorithm (previously developed [43]) at the end of each simulation run, corresponding to the end of the year 2017, for identification of clusters among persons with HIV diagnosed between 2015 and 2017 [19]. The algorithm identifies clusters using time as a proxy for genetic distance. Specifically, we approximated genetic distance between any two nodes as the sum of the difference between each node's time of diagnosis and time of infection of the node that commonly connects them in a transmission network. Assuming that the viral sequence will change approximately 1%, on average, over a 10-year interval, we replicated a genetic distance threshold of 0.5% in the model by defining clusters as any group of nodes connected in a single component within a 60-month time interval. Further, we classified clusters that include at least one person with HIV diagnosed in the most recent year (2017) as priority clusters. Because the completeness of sequence data (the proportion of all diagnosed infections for which a sequence is available) will impact cluster detection, we replicated incomplete data in the simulation: to replicate 48% sequence completeness in NHSS. For each run, we randomly selected 52% of infections diagnosed during 2015 to 2017 and excluded them in the generation of clusters.

We compared simulated clusters with molecular clusters, including the distribution of diagnosed cases by cluster type and cluster size and the distribution of number of clusters of varying size (Figure 7). We see in the figure that NHSS results are within the range of model results for each of these metrics. The proportion of persons with diagnosed infection that are part of a cluster as well as the distribution of persons by size of cluster in the model is consistent with NHSS results (plot on top in Figure 7). The distribution of clusters by size in the model is also consistent with the NHSS results (bottom plot in Figure 7). These results thus validate the methods in the model leading to the formation of the contact network structure and transmissions. Further, the model estimates also match well for the proportion of diagnosed cases in priority clusters, i.e., clusters with higher number of cases diagnosed in recent years, suggesting that the model generated events of diagnosis and care are similar to that observed in surveillance data (plot on top in Figure 7).

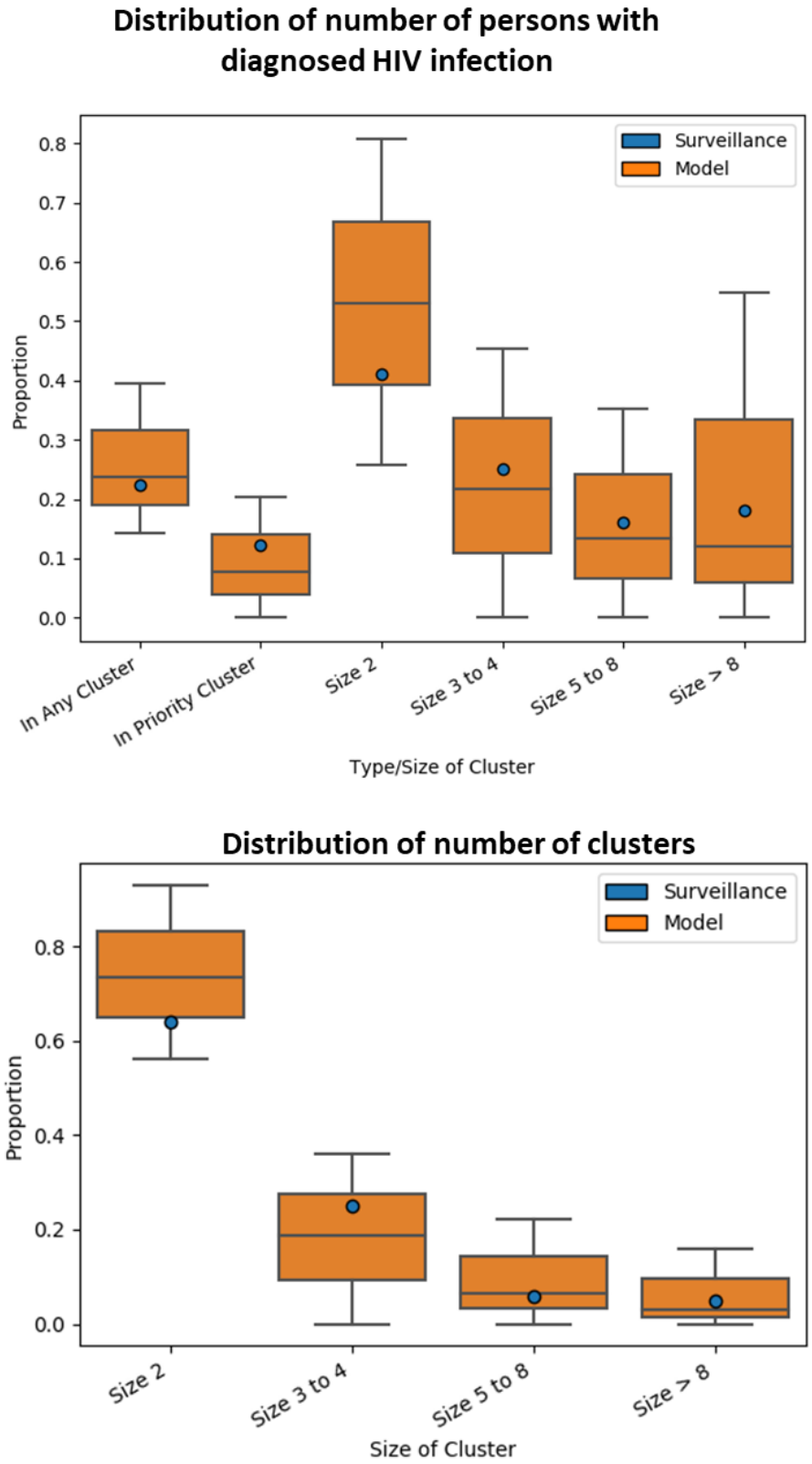

Figure 7. Distributions of number of persons diagnosed in years 2015, 2016, or 2017, by cluster type and size (top) and number of clusters by cluster size (bottom).

5. Discussion and conclusions

This paper presents a newly developed simulation method, ABENM with ECNA, for simulation of infectious disease epidemic projections for diseases with low prevalence, i.e., diseases with a small ratio of the size of infected population to size of susceptible population. We used ABENM for developing the PATH 4.0, a model for simulating HIV epidemic projections, and applied it for simulating HIV in the United States. PATH 4.0 is a comprehensive stochastic simulation model that simulates multiple interacting HIV-related events at the individual-level, including those related to sexual behavior, HIV-related care behavior, and disease progression.

Model estimates from PATH 4.0 on multiple surveillance outcomes related to both HIV disease projections and transmission network dynamics were comparable with surveillance data, thus validating the newly proposed ABENM simulation method and ECNA for network generation. Further, the fact that the clusters generated using the model compare well with those identified using HIV molecular sequence data supports the potential use of this model for analysis of cluster-based interventions.

However, there are limitations to this model. PATH 4.0 only simulated sexual transmissions of HIV, and thus, clusters formed through needle sharing contacts are not included in our results. About 9% of new HIV diagnoses are among PWID [62]. Future work could expand PATH 4.0 to include PWID for analyses of interventions specific to mode of transmission. We did not simulate the use of pre-exposure prophylaxis (PrEP) among uninfected persons, which can reduce the transmission rate. PrEP coverage is a recently introduced metric in NHSS reporting, available for year starting 2017 (about 12.5% in 2017) [63] and thus, analyses of clusters for future years should consider adding PrEP into the simulation. This can be done through adding a PrEP status to every susceptible node (as those who are a contact of an infected node are agents in the simulation) and incorporating the effectiveness of that into the Bernoulli transmission equation. We only modeled hypothetical jurisdictions to generate network dynamics, specifically to generate sub-networks with small amounts of mixing across sub-networks, as it is more likely that partnerships form between people within a certain geographic proximity. The jurisdictions are hypothetical in that we did not simulate jurisdiction-specific demographics or epidemic features. Despite these limitations, we believe the model presented here can help study questions specific to inform HIV cluster-based interventions, and the methods presented here could be utilized for construction of more sub-group specific models, such as by geographical jurisdiction or mode of transmission.

In summary, in this paper, we present the newly developed ABENM simulation method for simulation of diseases with low prevalence and its application to the development of the PATH 4.0 model. We propose PATH 4.0 as a tool for studying questions related to understanding the dynamics of cluster growth and its eventual application to identification of suitable cluster detection and intervention strategies. Cluster detection is a key component of the U.S. *Ending the HIV Epidemic* [20] national strategic plan. While surveillance is critical for the detection of clusters, a model in conjunction with surveillance can be used to refine cluster detection methods, better understand factors associated with cluster growth, and assess interventions to inform effective response strategies for prevention. As surveillance data are only available for cases that are diagnosed and reported, a model is a critical tool for understanding the true size of the clusters and assess key questions, such as the relative contributions of clusters to onward transmissions. As per our knowledge, this is the first model to have successfully replicated cluster features similar to that observed in molecular analysis of NHSS

data. Thus, PATH 4.0 serves as a novel tool for assessing intervention strategies for cluster detection and response. Moreover, the fact that this model more closely approximates true HIV transmission dynamics, including clusters of transmission occurring as a result of the scale-free network and nonrandom mixing, indicates that it would be a stronger mechanism, than an agent-based model alone, for making inferences about the benefits of various approaches to deploying HIV prevention interventions even in non-cluster response settings.

The successful replication of HIV disease patterns and cluster formation supports exploring the use of ABENM and ECNA in other areas. ABENM is specifically suited for diseases where simulation of contact networks is an essential component for accurate epidemic projections, and where the diseases have low prevalence that makes current network modeling methods challenging to use. In addition to HIV, other infectious diseases that fall under this category include tuberculosis, and chronic hepatitis B and C, or reemerging disease outbreaks such as SARS (severe acute respiratory syndrome), MERS (Middle East respiratory syndrome), and Ebola disease. In these cases, infection spreads through close contact between people, making the modeling of contact networks essential. These diseases also have high mortality and economic burdens, such that, in the case of remerging disease outbreaks that spread quickly, halting the disease in the very early stages of the outbreak (i.e., when the prevalence is low) is key for effective control.

Disclaimer

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Funding

SS and CG were funded 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.

Acknowledgments

We would like to acknowledge Dr. Timothy Green from the Centers for Disease Control and Prevention for his inputs in manuscript preparation.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. *Centers for Disease Control and Prevention. Estimated HIV incidence and prevalence in the United States, 2014–2018*. HIV Surveillance Supplemental Report 2020; 25(No. 1). Available from: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published May 2020. Accessed July 2020.

2. *Centers for Disease Control and Prevention*. HIV Surveillance Report, 2018 (Updated); vol.31. Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-report-2018-updated-vol-31.pdf. Published May 2020. Accessed September 2020.
3. P. G. Farnham, D. R. Holtgrave, C. Gopalappa, A. B. Hutchinson, S. L. Sansom, Lifetime costs and quality-adjusted life years saved from HIV prevention in the test and treat era, *J. Acquir. Immune Defic. Syndr.*, **64** (2013), e15–18. doi: 10.1097/QAI.0b013e3182a5c8d4
4. N. S. Harris, A. S. Johnson, Y. A. Huang, D. Kern, P. Fulton, D. K. Smith, et al., Vital signs: Status of human immunodeficiency virus testing, viral suppression, and HIV preexposure prophylaxis—United States, 2013–2018. *MMWR Morb. Mortal Wkly. Rep.*, **68** (2019), 1117–1123.
5. *Centers for Disease Control and Prevention.* Effectiveness of Prevention Strategies to Reduce the Risk of Acquiring or Transmitting HIV. Available from: https://www.cdc.gov/hiv/risk/estimates/preventionstrategies.html Accessed May 2020.
6. C. Gopalappa, P. G. Farnham, Y. H. Chen, S. L. Sansom, Progression and Transmission of HIV/AIDS (PATH 2.0), *Med. Decis. Making.*, **37** (2017), 224–233. doi: 10.1177/0272989X16668509
7. 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. doi: 10.1097/QAI.0000000000001707
8. H. W. Hethcote, J. W. Van Ark, Modeling HIV transmission and AIDS in the United States, *Springer Sci. Business Media*, **95** (2013).
9. Z. Li, D. W. Purcell, S. L. Sansom, D. Hayes, H. I. Hall, Vital signs: HIV transmission along the continuum of care - United States, 2016. *MMWR Morb. Mortal Wkly. Rep.*, **68** (2019), 267–272.
10. E. F. Long, M. L. Brandeau, D. K. Owens, The cost-effectiveness and population outcomes of expanded HIV screening and antiretroviral treatment in the United States, *Ann. Intern. Med.*, **153** (2010), 778–789. doi: 10.7326/0003-4819-153-12-201012210-00004
11. J. A. Jacquez, C. P. Simon, J. Koopman, L. Sattenspiel, T. Perry, Modeling and analyzing HIV transmission: The effect of contact patterns, *Math. Biosci.*, **92** (1988). doi: 10.1016/0025-5564(88)90031-4
12. E. A. Hernandez-Vargas, R. H. Middleton, Modeling the three stages in HIV infection, *J. Theor. Biol.*, **320** (2013), 33–40. doi: 10.1016/j.jtbi.2012.11.028
13. A. Lasry, S. L. Sansom, K. A. Hicks, V. Uzunangelov, A model for allocating CDC's HIV prevention resources in the United States, *Health Care Manag. Sci.*, **14** (2011), 115–124. doi: 10.1007/s10729-010-9147-2
14. S. W. Sorensen, S. L. Sansom, J. T. Brooks, G. Marks, E. M. Begier, K. Buchacz, et al., A mathematical model of comprehensive test-and-treat services and HIV incidence among men who have sex with men in the United States, *PloS One*, **7** (2012), e29098. doi: 10.1371/journal.pone.0029098
15. D. R. Holtgrave, Development of year 2020 goals for the National HIV/AIDS Strategy for the United States, *AIDS Behav.*, **18** (2014), 638–643. doi: 10.1007/s10461-013-0579-9
16. B. M. Adams, H. T. Banks, M. Davidian, H-D. Kwon, H. T. Tran, S. N. Wynne, et al., HIV dynamics: modeling, data analysis, and optimal treatment protocols, *J. Comput. Appl. Math.*, **184** (2005), 10–49.

17. E. Uzun 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.
18. A. M. Oster, A. M. France, J. Mermin, Molecular epidemiology and the transformation of HIV prevention, *JAMA*, **319** (2018), 1657–1658. doi: 10.1001/jama.2018.1513
19. A. M. Oster, A. M. France, N. Panneer, M. C. Bañez Ocfemia, E. Campbell, S. Dasgupta, et al., Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data, *J. Acquir. Immune Defic. Syndr.*, **79** (2018), 543–550. doi: 10.1097/QAI.0000000000001856
20. A. S. Fauci, R. R. Redfield, G. Sigounas, M. D. Weahkee, B. P. Giroir, Ending the HIV epidemic: A plan for the United States, *JAMA*, **321** (2019), 844–845.
21. R. Chou, C. Evans, A. Hoverman, C. Sun, T. Dana, C. Bougatsos, et al., Preexposure prophylaxis for the prevention of HIV infection, *JAMA,* **321 (**2019), 2214–2230. doi: 10.1001/jama.2019.2591
22. J. M. Baeten, D. Donnell, P. Ndase, N. R. Mugo, J. D. Campbell, J. Wangisi, et al., Antiretroviral prophylaxis for HIV prevention in heterosexual men and women, *N. Engl. J. Med*., **367** (2012), 399–410. doi: 10.1056/NEJMoa1108524
23. R. M. Grant, J. R. Lama, P. L. Anderson, V. McMahan, A. Y. Liu, L. Vargas, et al., Preexposure chemoprophylaxis for HIV prevention in men who have sex with men, *N. Engl. J. Med*., **363** (2010), 2587–2599. doi: 10.1056/NEJMoa1011205
24. M. C. Thigpen, P. M. Kebaabetswe, L. A. Paxton, D. K. Smith, S. R. Pathak, F. A. Soud, et al., Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana, *N. Engl. J. Med*., **367** (2012), 423–434. doi: 10.1056/NEJMoa1110711
25. S. L. Sansom, K. A. Hicks, J. Carrico, E. U. Jacobson, R. K. Shrestha, T. A. Green, et al., Optimal allocation of docietal HIV prevention resources to reduce HIV incidence in the United States, *Am. J. Public Health*., **111** (2021), 150–158. doi: 10.2105/AJPH.2020.305965
26. D. R. Gibson, N. M. Flynn, D. Perales, Effectiveness of syringe exchange programs in reducingHIV risk behavior and HIV seroconversion among injecting drug users, *AIDS*., **15** (2001), 1329–1341. doi: 10.1097/00002030-200107270-00002
27. R. M. Fernandes, M. Cary, G. Duarte, G. Jesus, J. Alarcão, C. Torre, et al., Effectiveness of needle and syringe Programmes in people who inject drugs—An overview of systematic reviews, *BMC Public Health,* **17** (2017), 309. doi: 10.1186/s12889-017-4210-2
28. M. Adams, Q. An, D. Broz, J. Burnett, C. Wejnert, G. Paz-Bailey, NHBS Study Group, Distributive syringe sharing and use of syringe services programs (SSPs) among persons who inject drugs, *AIDS Behav*., **23 (**2019), 3306–3314. doi: 10.1007/s10461-019-02615-4
29. E. J. Aspinall, D. Nambiar, D. J. Goldberg, M. Hickman, A. Weir, E. Van Velzen, et al., Are needle and syringe programmes associated with a reduction in HIV transmission among people who inject drugs: A systematic review and meta-analysis, *Int. J. Epidemiol*., **43** (2014), 235–248. doi: 10.1093/ije/dyt243
30. D. des Jarlais, A. Nugent, A. Solberg, J. Feelemyer, J. Mermin, D. Holtzman, Syringe service programs for persons who inject drugs in urban, suburban, and rural areas—United States, 2013, *MMWR Morb. Mortal Wkly. Rep*., **64** (2015), 1337–1341.

31. M. Eden, R. Castonguay, B. Munkhbat, H. Balasubramanian, C. Gopalappa., Agent-based evolving network modeling: A new simulation method for modeling diseases with low prevalence, *Health Care Manag. Sci.*, (2019). In Press.
32. C. I. Siettos, L. Russo, Mathematical modeling of infectious disease dynamics, *Virulence*, **4** (2013), 295–306. doi: 10.4161/viru.24041
33. T. Smieszek, L. Fiebig, R. W. Scholz, Models of epidemics: When contact repetition and clustering should be included, *Theor. Biol. Med. Model.*, **6** (2009), 11. doi: 10.1186/1742-4682-6-11
34. A-L. Barabasi, R. Albert, Emergence of scaling in random networks, *Science (New York, NY)*, **286** (1999), 509–512. doi: 10.1126/science.286.5439.509
35. A. M. El-Sayed, P. Scarborough, L. Seemann, S. Galea, Social network analysis and agent-based modeling in social epidemiology, *Epidemiol. Perspect. Innov.*, **9** (2012), 1. doi: 10.1186/1742-5573-9-1
36. *Centers for Disease Control and Prevention. Estimated HIV incidence and prevalence in the United States, 2010–2015.* HIV Surveillance Supplemental Report 2018; 23(No. 1). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-23-1.pdf Published March 2018. Accessed May 2020.
37. A. Lansky, T. Finlayson, C. Johnson, D. Holtzman, C. Wejnert, A. Mitsch, et al., Estimating the number of persons who inject drugs in the united states by meta-analysis to calculate national rates of HIV and hepatitis C virus infections, *PloS One*, **9** (2014), e97596. doi: 10.1371/journal.pone.0097596
38. D. W. Purcell, C. H. Johnson, A. Lansky, J. Prejean, R. Stein, P. Denning, et al., Estimating the population size of men who have sex with men in the United States to obtain HIV and syphilis rates, *Open AIDS J.*, **6** (2012), 98–107. doi: 10.2174/1874613601206010098
39. F. Liljeros, C. R. Edling, L. A. Amaral, H. E. Stanley, Y. Åberg, The web of human sexual contacts, *Nature*, **411** (2001), 907–908. doi: 10.1038/35082140
40. J. O. Wertheim, S. L. Kosakovsky Pond, L. A. Forgione, S. R. Mehta, B. Murrell, S. Shah, et al., Social and genetic networks of HIV-1 transmission in New York City, *PloS Pathog*., **13** (2017), e1006000. doi: 10.1371/journal.ppat.1006000
41. J. O Wertheim, A. J. Leigh Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D.M. Smith, et al., The global transmission network of HIV-1, *J. Infect. Dis.*, **209** (2014), 304–313. doi: 10.1093/infdis/jit524
42. B. Fotouhi, M. G. Rabbat, Degree correlation in scale-free graphs, *Eur. Phys. J. B.*, **85** (2013), 510.
43. Y. H. Chen, A. M. France, P. G. Farnham, S. L. Sansom, C. Gopalappa, A. Oster., Replicating HIV Transmission Clusters in a U.S. HIV Agent-Based Model [abstract]. In: Abstracts: SMDM 40th Annual Meeting; 2018 Oct; Montréal, Québec, Canada.
44. U. Wilensky, NetLogo. [Internet]. 1999. Available from: http://ccl.northwestern.edu/netlogo/.
45. A. R. Board, L. Linley, A. M. Oster, M. Watson, R. Song, T. Zhang, et al., Geographic distribution of HIV transmission networks in the United States, *J. Acquir. Immune Defic. Syndr.*, (2020). In Press.
46. K. Buchacz, C. Armon, F. J. Palella, R. K. Baker, E. Tedaldi, M. D. Durham, et al., CD4 cell counts at HIV diagnosis among HIV outpatient study participants, 2000–2009, *AIDS Res. Treat.*, **2012** (2012), 869841.

47. T. J. Finlayson, B. Le, A. Smith, K. Bowles, M. Cribbin, I. Miles, et al., Centers for disease control and prevention (CDC), HIV risk, prevention, and testing behaviors among men who have sex with men—National HIV Behavioral Surveillance System, 21 U.S. cities, United States, 2008, *MMWR Surveill. Summ.,* **60** (2011), 1–34.
48. A. Chandra, W. D. Mosher, C. Copen, C. Sionean, Sexual behavior, sexual attraction, and sexual identity in the United States: Data from the 2006–2008 National Survey of Family Growth, *Natl. Health Stat. Report.,* **36** (2011), 1–36.
49. J. A. Grey, K. T. Bernstein, P. S. Sullivan, D. W. Purcell, H. W. Chesson, T. L. Gift, et al., Estimating the population sizes of men who have sex with men in US states and counties using data from the American Community Survey, *JMIR Public Health Surveill.*, **2** (2016), e14.
50. M. Reece, D. Herbenick, V. Schick, S. A. Sanders, B. Dodge, J. D. Fortenberry, Background and considerations on the national survey of sexual health and behavior (NSSHB) from the investigators, *J. Sex. Med.*, **7** (2010), 243–245. doi: 10.1111/j.1743-6109.2010.02038.x
51. S. M. Cohen, K. M. Gray, M. C. Ocfemia, A. S. Johnson, H. I. Hall, The status of the national HIV surveillance system, United States, 2013, *Public Health Rep.*, **129** (2014), 335–341. doi: 10.1177/003335491412900408
52. *Centers for Disease Control and Prevention. Behavioral and Clinical Characteristics of Persons with Diagnosed HIV Infection—Medical Monitoring Project, United States, 2016 Cycle (June 2016–May 2017).* HIV Surveillance Special Report 21. Revised edition. Available from: https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published June 2019. Accessed Feb 2021.
53. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas—2011.* HIV Surveillance Supplemental Report 2013; 18(No. 5). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-18-5.pdf Published October 2013. Accessed May 2020.
54. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas—2012.* HIV Surveillance Supplemental Report 2014; 19(No. 3). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-19-3.pdf Published November 2014. Accessed May 2020.
55. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas, 2015.* HIV Surveillance Supplemental Report 2017; 22(No. 2). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-22-2.pdf Published July 2017. Accessed May 2020.
56. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas, 2014.* HIV Surveillance Supplemental Report 2016; 21(No. 4). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-21-4.pdf Published July 2016. Accessed May 2020.
57. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas—2013.* HIV

Surveillance Supplemental Report 2015; 20(No. 2). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-20-2.pdf Published July 2015. Accessed May 2020.

58. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas, 2016.* HIV Surveillance Supplemental Report 2018; 23(No. 4). Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-23-4.pdf Published June 2018. Accessed May 2020.
59. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas, 2018.* HIV Surveillance Supplemental Report 2020; 25(No. 2). Available from: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published May 2020. Accessed July 2020.
60. *Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas, 2017.* HIV Surveillance Supplemental Report 2019; 24(No. 3). Available from: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published June 2019. Accessed July 2020.
61. S. L. K. Pond, S. Weaver, A. J. L. Brown, J. O. Wertheim, HIV-TRACE (TRAnsmission Cluster Engine): A tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens, *Mol. Biol. Evol.*, **35** (2018), 1812–1819. doi: 10.1093/molbev/msy016
62. *Centers for Disease Control and Prevention.* HIV Surveillance Report, 2017; vol. 29. Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-report-2017-vol-29.pdf. Published November 2018. Accessed September 2020.
63. *Centers for Disease Control and Prevention.* NCHHSTP AtlasPlus. Updated 2019. Available from: https://gis.cdc.gov/grasp/nchhstpatlas/tables.html. Accessed July 2020.

AIMS Press