SimpactCyan 1.0: An Open-source Simulator for Individual-Based Models in HIV Epidemiology with R and Python Interfaces

Jori Liesenborgs[1], **Diana M Hendrickx**[2], **Elise Kuylen**[3,4], **David Niyukuri**[5,6], **Niel Hens**[2,4], **and Wim Delva**[2,5,6,7,8,9,*]

[1]Expertise Centre for Digital Media, Hasselt University – tUL, Diepenbeek, Belgium
[2]Center for Statistics, I-BioStat, Hasselt University, Diepenbeek, Belgium
[3]IDLab, University of Antwerp, Antwerp, Belgium
[4]Centre for Health Economics Research and Modelling Infectious Diseases and Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, Belgium
[5]The South African Department of Science and Technology-National Research Foundation (DST-NRF) Centre of Excellence in Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, Stellenbosch, South Africa
[6]Department of Global Health, Faculty of Medicine and Health, Stellenbosch University, Stellenbosch, South Africa
[7]International Centre for Reproductive Health, Ghent University, Ghent, Belgium
[8]Rega Institute for Medical Research, KU Leuven, Leuven, Belgium
[9]School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa
[*]wimdelva@gmail.com

ABSTRACT

SimpactCyan is an open-source simulator for individual-based models in HIV epidemiology. Its core algorithm is written in C++ for computational efficiency, while the R and Python interfaces aim to make the tool accessible to the fast-growing community of R and Python users. Transmission, treatment and prevention of HIV infections in dynamic sexual networks are simulated by discrete events. A generic "intervention" event allows model parameters to be changed over time, and can be used to model medical and behavioural HIV prevention programmes. First, we describe a more efficient variant of the modified Next Reaction Method that drives our continuous-time simulator. Next, we outline key built-in features and assumptions of individual-based models formulated in SimpactCyan, and provide code snippets for how to formulate, execute and analyse models in SimpactCyan through its R and Python interfaces. Lastly, we give two examples of applications in HIV epidemiology: the first demonstrates how the software can be used to estimate the impact of progressive changes to the eligibility criteria for HIV treatment on HIV incidence. The second example illustrates the use of SimpactCyan as a data-generating tool for assessing the performance of a phylodynamic inference framework.

Introduction

In epidemiology, mathematical models are widely used to simulate progression, transmission, prevention and treatment of infectious diseases. The majority of these models are deterministic compartmental models, simulating population averages of changes in infection status and disease stages over time. However, many infectious diseases, in particular sexually transmitted diseases, are subject to high individual heterogeneity. Unlike compartmental models simulating population averages, individual-based models (IBMs) keep track of the events that happen to each individual separately, and are therefore able to take into account various sources of individual heterogeneity[1].

The ability to let population-level features of complex systems emerge from processes and events that happen to interacting individuals, is arguably the most important quality of IBMs. As the computational expense of IBMs has become less prohibitive thanks to multi-core processors and increased access to high-performance computers, there is a growing use of IBMs in infectious disease epidemiology[2]. SimpactCyan is conceived as a versatile model-building tool to address research questions in HIV epidemiology at the intersection of network and social epidemiology, computational biology, public health and policy modelling.

A large amount of general frameworks for individual-based simulations have been developed in the last decades. These platforms vary widely in terms of platform properties, usability, operating ability, pragmatics and security management, which

Tool	Time implementation	Sexual network	R interface	Python interface	Source code available online	Reference
SimpactCyan	continuous, mNRM	dynamic	✓	✓	✓	[5–7]
GEMFsim	continuous, Gillespie algorithm	static	✓	✓	✓	[8]
FAVITES	continuous, Gillespie algorithm	static	✓	✓	✓	[9]
EpiModel	discrete	dynamic	✓	✗	✓	[10]
STDsim	discrete	dynamic	✗	✗	✗	[11]
NetLogo	discrete	dynamic	✓	✗	✓	[12, 13]
EMOD	discrete	dynamic	✗	✗	✓	[14]
HIV-CDM	discrete	dynamic	✗	✗	✗	[15]
MicroCOSM	discrete	dynamic	✗	✗	✗	[16]
Path 2.0	discrete	dynamic	✓ (through RNetLogo)	✗	✗	[17]
TITAN	discrete	dynamic	✓	✓	✓	[18]

Table 1. Functional and structural differences between SimpactCyan and existing tools for IBM studies in HIV epidemiology.

makes it difficult to choose the most suitable framework for simulation in the context of a particular research question[3].

Table 1 summarises functional and structural differences between SimpactCyan and ten other tools for constructing individual-based models of HIV transmission between individuals connected via sexual relationships. These tools were identified by an ongoing systematic review of simulation-based methods for the calibration of individual-based models to summary data in epidemiology[4].

Current software for implementing IBMs to address questions in HIV epidemiology has several shortcomings. While some modelling tools (e.g. STDSIM for simulating transmission of HIV and other Sexually Transmitted Diseases[11]) are not open source, other IBMs (e.g. EMOD[14]) are relatively difficult to modify. Another limitation of EMOD is that it can only be used on computers running Windows 10, Windows Server 12, Windows HPC Server 12 or CentOS 7.1. Furthermore, while it has interfaces for Matlab and Python, it does not have an R interface. NetLogo models, on the other hand, are easily modifiable[3], and can be run from within the R environment[12], but are prohibitively slow for simulating large populations over the time-scale relevant for HIV epidemiology.

With a few exceptions (the GEMFsim[8] and FAVITES[9] simulators), existing simulators that can be readily used in HIV epidemiology implement IBMs in discrete time rather than continuous time. By simulation in continuous time we mean that events can take place and subsequently the state of the system can be updated at any point in time. It also means that the time interval between the execution of two consecutive events is only limited by the numeric precision of the implementation of the method used to sample event times. A continuous time implementation of IBMs has the advantage that it elegantly handles competing risks to multiple events. For instance, an individual may be concurrently at risk of HIV-related mortality and at risk of transmitting the virus to a partner. Evaluating the model in fixed time steps may lead to the situation where both events are scheduled to have taken place between now and the next time step. However, in reality, this is only possible if the transmission event happens first. In the continuous time model evaluation, we know exactly which of the two events is scheduled first, and logical consequences for the likelihood of subsequent events are processed along with the execution of the first event. Furthermore, events happening after short and long time periods can be included in a single simulation. In contrast, in a discrete time model, simulating events that occur on vastly different time-scales can be computationally inefficient. Frequently occurring events may require a small time step, possibly leading to the occurrence of rare events being evaluated with a much higher frequency than necessary.

Another limitation of existing implementations of IBMs for dynamic sexual networks, is that they consider possible relationships sequentially. As a consequence they require ad-hoc assumptions and decisions about in what order people “go out” to find partners and can “be found”. For example, in EMOD, males and females are placed in a separate queue, where they stay for a predefined period, after which they form a relationship[14]. In STDsim, the order of going out to search for a partner and for being available for a relationship are both determined by stochastic processes[11].

SimpactCyan is a simulator for event-driven IBMs in HIV epidemiology, evaluated in continuous time: the state of the system is updated each time an event happens. Furthermore, all possible relationships are considered simultaneously instead of sequentially.

Simpact (*SimpactWhite*) was first developed in Matlab[19–21]. Later, variants were developed as a MASON Multi-agent Simulation Toolkit in Java (*SimpactBlue*), and in Python (*SimpactyPurple*)[22]. To improve both speed and user-friendliness of the tool, we embarked on a major overhaul in 2013, leading to the current version (*SimpactCyan*) that combines a computationally efficient simulation engine written in C++ with R and Python interfaces. An exhaustive, deep comparison of SimpactCyan with all prior Simpact programs is beyond the scope of this paper, and arguably not crucially important, for the simple reason that there is no ongoing development of nor support for any of the legacy versions (SimpactWhite, SimpactBlue and SimpactPurple) and these versions are no longer in use. However, it may be useful to give some perspective of the relative improvement. In the early stages of SimpactCyan development, we conducted a comparison study that indicated runtimes for SimpactCyan were up to 280 times shorter, compared to SimpactWhite[23].

In this paper, we describe a more efficient variant of the modified Next Reaction Method (mNRM) that drives the simulator, we outline key built-in features and assumptions of individual-based models formulated in SimpactCyan, and provide code snippets for how to formulate, execute and analyse models in SimpactCyan through its R and Python interfaces. As runtimes for SimpactCyan strongly depend on population size and the intensity with which relationships are formed and dissolved in the population, we present the results of a concise simulation study to provide additional insights and visual representation of these associations. We end by giving two examples of applications in HIV epidemiology: the first demonstrates how the software can be used to estimate the impact that changes to the eligibility criteria for antiretroviral therapy (ART) had on HIV incidence in a hyperendemic setting. The second example illustrates the use of SimpactCyan as a data-generating tool for assessing the performance of other modelling frameworks.

Discrete events simulation algorithm

The modified Next Reaction Method (mNRM)

Event times, i.e. time points in the simulation at which events are scheduled to take place, are determined using the *modified Next Reaction Method* (mNRM)[24], a more efficient variant of the Gillespie algorithm[25–27] and the Next Reaction Method[28]. The mNRM was originally designed for simulating chemical systems with time-dependent propensities and delays, but in SimpactCyan we use it to simulate how individuals are at risk of events according to time-dependent hazard functions. In the mNRM algorithm, there is a core distinction between *internal event times* and (simulated) *real-world event times*. The internal event times determine when an event will be triggered according to the event's *internal clock*. Internal clock time advances faster as the hazard for the event increases. By real-world time we mean the calendar time in the simulated population.

Calling the *internal* time interval until a specific event fires ΔT, such internal time intervals are randomly sampled from an exponential distribution: $\Delta T \sim \mathrm{Exp}(1)$.

The event's *hazard function* $h(\bullet)$, referred to as the *propensity function* in[24], maps the internal time interval ΔT until the event fires onto Δt, a real-world time interval,

$$\Delta T = \int_t^{t+\Delta t} h(X(t),s)ds, \tag{1}$$

in which t is the previous time an event was triggered. (This corresponds to equation 13 in the original article[24], where we have omitted the event-specific index for concision.) It is this hazard h that can depend on the state $X(t)$ of the simulation, and possibly also explicitly on time t. In SimpactCyan, the state of the simulation is made up of all the individuals in the population and their respective properties, such as their age, gender, HIV infection status, ART status, and whom they are in relationships with. This state $X(t)$ does *not* depend on time in a *continuous* manner, it only changes when an event is fired, i.e. when its internal time interval expires. Note that the formula above is for a single event, and while ΔT itself is not affected by other events, the mapping onto Δt certainly can be: other events can change the simulation state, and the hazard of the event depends on this state.

The main idea is illustrated in Figure 1: internal time intervals are chosen from an exponential distribution, and are mapped onto real-world time intervals through hazard functions. Because hazards can depend on the simulation state and can have an explicit time dependency, this mapping can be rather complex.

While the hazard *can* cause complex behaviour, this is of course not necessarily the case. If one uses a constant hazard, this merely causes a linear scaling between internal time ΔT and real-world time Δt:

$$\Delta T = h\Delta t \quad \text{(for a constant hazard).}$$

This also illustrates that the larger the hazard, the earlier the event will fire, i.e. the real-world time interval will be smaller.

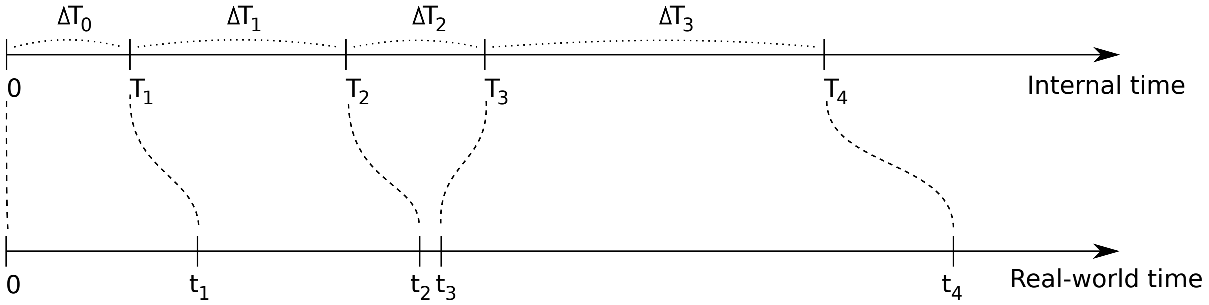

Figure 1. In the modified Next Reaction Method, intervals ΔT_i (in general for different events) are generated independently from other events in a straightforward manner, using an exponential probability distribution ($\Delta T_i \sim \mathrm{Exp}(1)$), and are used to advance an *internal* clock T. Using the notion of a hazard function (1), these internal time intervals are mapped onto intervals Δt_i, which advance a (simulated) *real-world* time t, and need not have a straightforward relation to the internal times: a small internal time difference can lead to a large real-world time difference and vice versa. It is through this hazard function that interdependencies between events can be introduced.

As an example, let's consider the event of forming a heterosexual relationship. At a certain time in the simulation, many formation events will be scheduled, one event for each man-woman pair that can possibly form a relationship. The internal time interval for each of these events will be drawn from the simple exponential distribution. The mapping to a real-world time at which the event will fire, is done using the hazard-based method, and the event that will take place next is the one that will have the smallest of these real-world times (cfr. the time ordering in step 6 of algorithm 3 in[24]). This hazard depends on aspects of the simulation state as defined by the hazard function for relationship formation: how many relationships the man and woman of the candidate couple are already engaged in, what the preferred age differences with their respective partners are, etc. One can also imagine an explicit time dependency in the hazard: e.g. the hazard of forming a relationship increases as the time period since the relationship became possible gets longer.

While most of the events in SimpactCyan are scheduled using the exponential distribution to generate values for internal ΔT, some events are scheduled directly in real-world time. An example of this is the scheduling of the HIV 'seeding' event, i.e. the timing of introducing HIV into the population. This alternative method could still be thought of as a special case of internal and real-world time mapping. This is because if ΔT is set to the actual real-world time interval until the event fires, and the hazard is set to $h = 1$, internal and real-world time intervals match.

More efficient mNRM algorithm

Each time an event is triggered, the state of the simulation changes. Because the hazard of any event can depend on this state, in the most general version of the mNRM algorithm, one would recalculate the real-world event times of all remaining events each time an event gets triggered: this ensures that the possibly changed state is taken into account. Always recalculating all event fire times is computationally very inefficient, however. Although the state may have been changed somewhat, this change may not be relevant for many of the event hazards in use. As a result, most updated real-world event times would be the same as before.

To avoid unnecessary recalculations of event times, SimpactCyan employs a variant of the mNRM algorithm, in which each individual is linked to a list of events that involve him or her, and events that involve multiple people will appear on the lists of all of these individuals. For example, a mortality event would be present in the list of only one individual, while a relationship formation event concerns two people and would therefore appear on two such lists. Figure 2 illustrates this idea. The lists are meant to keep track of event times that may require recalculation as a result of another event firing. If the HIV-positive partner of Person X (not depicted in Figure 2) dies, this death will trigger several updates to the system, including that the deceased gets removed from the population, and dissolution events are triggered immediately (i.e. no recalculation needed) for all the relationships that the deceased was engaged in. Such dissolution events will subsequently lead to the removal of the HIV transmission event that was on Person X's list (also without any recalculation required).

When an event fires, only the properties of a very limited set of people are changed, hence one only needs to recalculate the fire times of the events in those people's lists. For example, when the event of Person A forming a relationship with Person B takes place, the real-world fire times for the events in the lists of Person A and Person B will be automatically recalculated. Apart from affecting the people in whose lists an event appears, some events can affect additional people. As an example, a birth event will only appear in the list of the pregnant woman and not in the event list of the father, because the scheduled birth should not be affected in the event of the death of the father. However, when triggered, the newborn will be listed as a child of the father. In general, the number of additionally affected people will be very small compared to the size of the population,

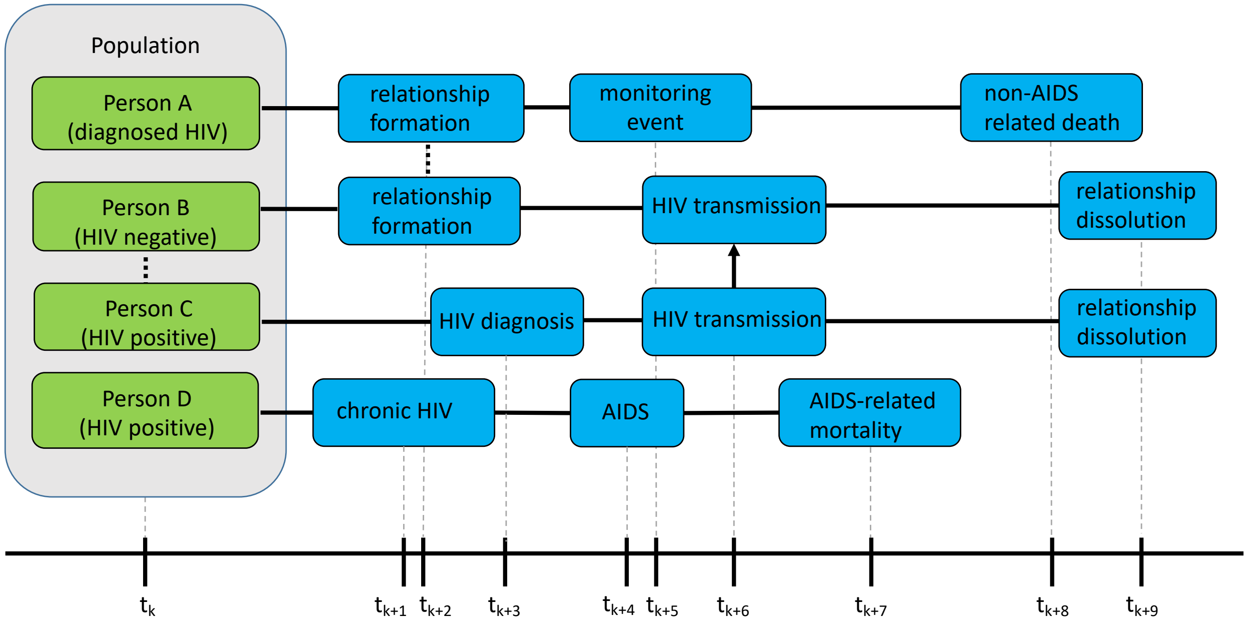

Figure 2. The Figure shows, for four different people in the population, what the next three scheduled events are after time point t_k. Solid lines represent time intervals; The dotted line represents the formation of a relationship; the arrow represents HIV transmission. We can see that Person A and Person B will form a relationship. Concurrent with their relationship to Person A, person B is also already in a relationship with Person C. We can see that an HIV transmission from Person C to Person B is scheduled, after Person C is diagnosed. After HIV transmission, person B and person C break up their relationship. After forming a relationship with Person B, Person A (who was already diagnosed with HIV) is monitored to follow up the progression of HIV. Person A dies a non-AIDS related death. Person D, who is HIV positive, will progress to the chronic stage of HIV infection, after which he or she will develop AIDS, and die of AIDS-related complications.

causing only a fraction of the event fire times to be recalculated. This allows the modified algorithm to run much faster than the basic algorithm that always recalculates all event times. Furthermore, fire times of events that are present in the event lists of two individuals (e.g. relationship formation), are recalculated by only one of them.

Besides these types of events, there are also ‘global’ events. These events do not refer to a particular person and will modify the state in a very general way. In general, when such a global event is triggered, this causes *all* other event fire times to be recalculated. Introducing HIV into the population through an HIV seeding event is an example of a global event.

Population and events in SimpactCyan

Model populations consist of men and/or women. They can be introduced into the simulation in two ways: (i) during the initialization of the simulation, in which case individuals with certain ages (drawn from a distribution) are added to the simulation, and (ii) through the birth of new individuals during the course of the simulation run.

Once born, an individual will become sexually active when a *debut event* is triggered. If the individual is introduced into the population at the start of the simulation, and the age exceeds the debut age, this event no longer needs to be scheduled. Every person always has a ‘normal’ *mortality event* scheduled, which corresponds to a cause of death other than AIDS.

To get an HIV epidemic started, an *HIV seeding event* must be scheduled. When this event is triggered, a number of people in the existing population will be marked as being HIV-infected. An infected individual will go through a number of infection stages, starting with acute HIV infection. After a default duration of 3 months[29], a *chronic stage event* is triggered, moving the individual to the chronic infection stage. A fixed amount of time before dying of AIDS (15 months by default)[29], an *AIDS stage event* is triggered, marking the transition of the chronic HIV stage to the AIDS stage. Six months before the expected AIDS-related death, a *final AIDS stage event* is triggered, after which the individual is in the ‘final AIDS stage’. It is assumed that one is too ill to be sexually active during this final stage[29]. When the *AIDS mortality event* is triggered, the individual dies of AIDS.

The `population.msm` parameter enables simulation of populations in which (some) men only form sexual relationships with other men, and/or can form relationships with with men and women. Under the default parameter setting (`population.msm = no`), every man-woman pair past the age of sexual debut can potentially form a relationship. For every such pair, a *formation event* is scheduled by sampling from the probability distribution that emerges as a result of the specified hazard function for relationship formation. An example for such a hazard function is the 'agegap' hazard, shown in the equation below. Only when an event of this type is triggered, an actual relationship is formed between the involved persons, which in turn can cause other events to get scheduled, e.g. a relationship dissolution event.

$$
\begin{aligned}
\text{hazard} &= F \times \exp\Bigg(\alpha_{\text{baseline}} + \alpha_{\text{numrel,man}} P_{\text{man}} + \alpha_{\text{numrel,woman}} P_{\text{woman}} \\
&+ \alpha_{\text{numrel,diff}} |P_{\text{man}} - P_{\text{woman}}| + \alpha_{\text{meanage}} (A_{\text{man}}(t) + A_{\text{woman}}(t))/2 \\
&+ \alpha_{\text{eagerness,sum}} (E_{\text{man}} + E_{\text{woman}}) + \alpha_{\text{eagerness,diff}} |E_{\text{man}} - E_{\text{woman}}| \\
&+ \alpha_{\text{dist}} |\vec{R}_{\text{man}} - \vec{R}_{\text{woman}}| \\
&+ \alpha_{\text{gap,factor,man}} |A_{\text{man}}(t) - A_{\text{woman}}(t) - D_{p,\text{man}} - \alpha_{\text{gap,agescale,man}} A_{\text{man}}(t)| \\
&+ \alpha_{\text{gap,factor,woman}} |A_{\text{man}}(t) - A_{\text{woman}}(t) - D_{p,\text{woman}} - \alpha_{\text{gap,agescale,woman}} A_{\text{woman}}(t)| \\
&+ \beta (t - t_{\text{ref}})\Bigg)
\end{aligned}
\tag{2}
$$

This complex looking hazard is actually of the form $\exp(A + Bt)$. The α parameters are weights that need to be set in the configuration of the simulation. They control the importance of various aspects of how individuals choose sexual partners. Via so-called intervention events, these weight parameters can be changed at arbitrary points in time during the simulation. In this way, temporal changes in sexual risk behaviours can be modelled, such as reductions in partner concurrency, age-disparate relationships or overall sexual activity levels. Variable P represents the number of relationships that an individual is currently engaged in, $A(t)$ the age of the individual, and E represents a person-specific sex drive, of which the distribution is user-defined to allow control over the amount of heterogeneity in sexual activity within the population. The effect of distance ($|\vec{R}_{\text{man}} - \vec{R}_{\text{woman}}|$) between the candidate partners can also be taken into account.

In the terms about the age gap, the age difference between the man and woman in the potential relationship is compared to the preferred age difference D_p (which defaults to zero). If the $\alpha_{\text{gap,agescale}}$ values are set to zero, then D_p will always be the preferred age gap; a positive value between zero and one will cause the preferred age difference to increase as the individual grows older. The $\alpha_{\text{gap,factor}}$ values at the start of these terms in turn determine the importance of the preferred age difference in the hazard. By setting the $\alpha_{\text{gap,factor}}$ to a negative value, the hazard decreases as the actual age difference between the candidate partners deviates from the preferred age difference.

The β parameter can be used to introduce an effect on the hazard that depends on the time since the relationship became possible. Here, t_{ref} refers to the point in time at which the relationship between the two candidate partners became possible. If no relationship existed between them earlier, this is the time at which the youngest person reached the debut age. On the other hand, if they were previously in a relationship with each other already, it is the time at which that relationship was dissolved. For completeness, the factor F is a normalization-like factor, to be able to use similar parameters for different population sizes.

A formation event results in the establishment of a sexual relationship, and subsequently, the female partner is at risk of falling pregnant. In that case a *conception event* will be triggered and a while later a *birth event* will take place, introducing a new individual into the population. In case one of the partners in the relationship is HIV-infected, transmission of the virus may occur. If so, a *transmission event* will fire, and the newly infected individual will go through the different infection stages as described earlier. Of course, it is also possible that the relationship will cease to exist, in which case a *dissolution event* will be triggered. Note that in the version at the time of writing, there is no mother-to-child-transmission (MTCT).

Starting ART and dropping out of treatment is managed by another sequence of events. When an individual becomes HIV-infected, either by HIV seeding or by transmission, first a *diagnosis event* is scheduled. Upon diagnosis, an *HIV monitoring event* is scheduled to monitor the progression of the HIV infection. When this event is fired, ART may be initiated, but only if the individual is both eligible (according to a CD4 cell count threshold) and willing to start HIV treatment; if not, a new monitoring event will be scheduled. If ART is initiated, no more monitoring events will be scheduled, but the individual will be at risk of discontinuing his or her HIV treatment, in which case a *dropout event* is triggered. When a person drops out of treatment, a new *diagnosis event* will be scheduled, which should be interpreted as an act of re-engagement in HIV Care[30].

Formulating, running and analysing IBMs with SimpactCyan from R or Python

Instructions for installing the core SimpactCyan program and its R interface (the Python interface is automatically installed along with the core program) can be found at `http://www.simpact.org/how-to-use-simpact/`. To set up a simulation, one needs to prepare a configuration file as a text file with key/value pairs, describing all parameters of the simulation, a snippet of which could look like this:

```
...
population.nummen     = 200
population.numwomen   = 200
population.simtime    = 40
...
```

Preparing the configuration file manually is time-consuming work however, as *all* event properties needed in a simulation must be set. To make it easier to prepare and run simulations, there is a Python module that can be used to control SimpactCyan from Python, or alternatively an R library that can be installed in R, with a similar interface. It is also possible to use a combined approach: first prepare a configuration file from within R or Python, and subsequently use this configuration to start simulations from the command-line. It can be very helpful when running simulations on a high performance computing cluster for example, where R or Python may not be available.

To use SimpactCyan from within an R session, the `RSimpactCyan` library must first be installed and loaded. This provides a `simpact.run` function that allows a simulation to be configured much more easily than using the configuration file mentioned above: instead of needing to set all parameters of a simulation, only the parameters that are different from the default values need to be specified. The full documentation of all the parameters that can be configured, what they mean and what their default values are, is found at https://simpactcyan.readthedocs.io/en/latest/simpact_simulationdetails.html. If only the key/value pairs in the code snippet above deviate from their default values, the configuration of the simulation would simply become:

```
cfg <- list()
cfg["population.nummen"] <- 200
cfg["population.numwomen"] <- 200
cfg["population.simtime"] <- 40
```

Similarly, the Python module `pysimpactcyan` defines a `PySimpactCyan` class with a `run` member function that also needs only the settings that differ from the defaults:

```
cfg = { }
cfg["population.nummen"] = 200
cfg["population.numwomen"] = 200
cfg["population.simtime"] = 40
```

Many of the configuration values will be character strings or numbers, but for some options it is allowed to specify one of the supported one- or two-dimensional probability distributions. For example, the `birth.pregnancyduration.dist.type` is by default set to `fixed` with a value corresponding to 268/365 (simulation times are expressed in years), such that every pregnant woman would give birth after precisely 268 days. To allow for some variability (e.g. a standard deviation of 16 days), a log-normal distribution could be used instead:

```
mu <- 268/365
var <- (16/365)^2
cfg["birth.pregnancyduration.dist.type"] <- "lognormal"
cfg["birth.pregnancyduration.dist.lognormal.zeta"] <- log(mu/sqrt(1+var/mu^2))
cfg["birth.pregnancyduration.dist.lognormal.sigma"] <- sqrt(log(1+var/mu^2))
```

Apart from using a fixed number, supported one-dimensional distributions are the beta, exponential, gamma, log-normal, normal and uniform distributions, as well as user-defined discrete distributions (e.g. based on the frequencies listed in a CSV file). For two-dimensional distributions, one can specify a fixed pair of values, or choose values from binormal or uniform distributions. Here too, user-defined discrete distributions can be specified.

Runtime analysis

We conducted a small simulation study with SimpactCyan to explore how the runtime varies as a function of population size (at the outset of the simulation), single-core versus multi-core execution of the simulation, fraction of the population of the opposite sex that each individual can potentially form relationships with, and number of relationships that are formed over the course of the simulation. Specifically, we designed scenarios of HIV epidemics in heterosexual populations and ran each scenario over a calendar time period of 40 years. Each scenario was repeated 10 times with different seeds for the random number generators, and the average runtime was calculated per scenario. Scenarios varied by initial population size (4 000, 10 000 and 20 000), the fraction of the population of the opposite sex that each individual could form relationships with (0.2 and 0.4), and the effect of the number of relationships already engaged in by candidate partners on the log-transformed hazard of these two individuals forming a relationship (i.e. the $\alpha_{\mathrm{numrel,man}}$ and $\alpha_{\mathrm{numrel,woman}}$ parameters in the hazard function (-0.5 and -5)). These parameters are strongly correlated with the number of relationships that are formed over the course of a simulation. Lastly, we executed simulations with the single-core and multi-core versions of the algorithm. Simulations were run on a MacBook Pro with Intel Core i9 6-core processor, running macOS Mojave (Version 10.14.3).

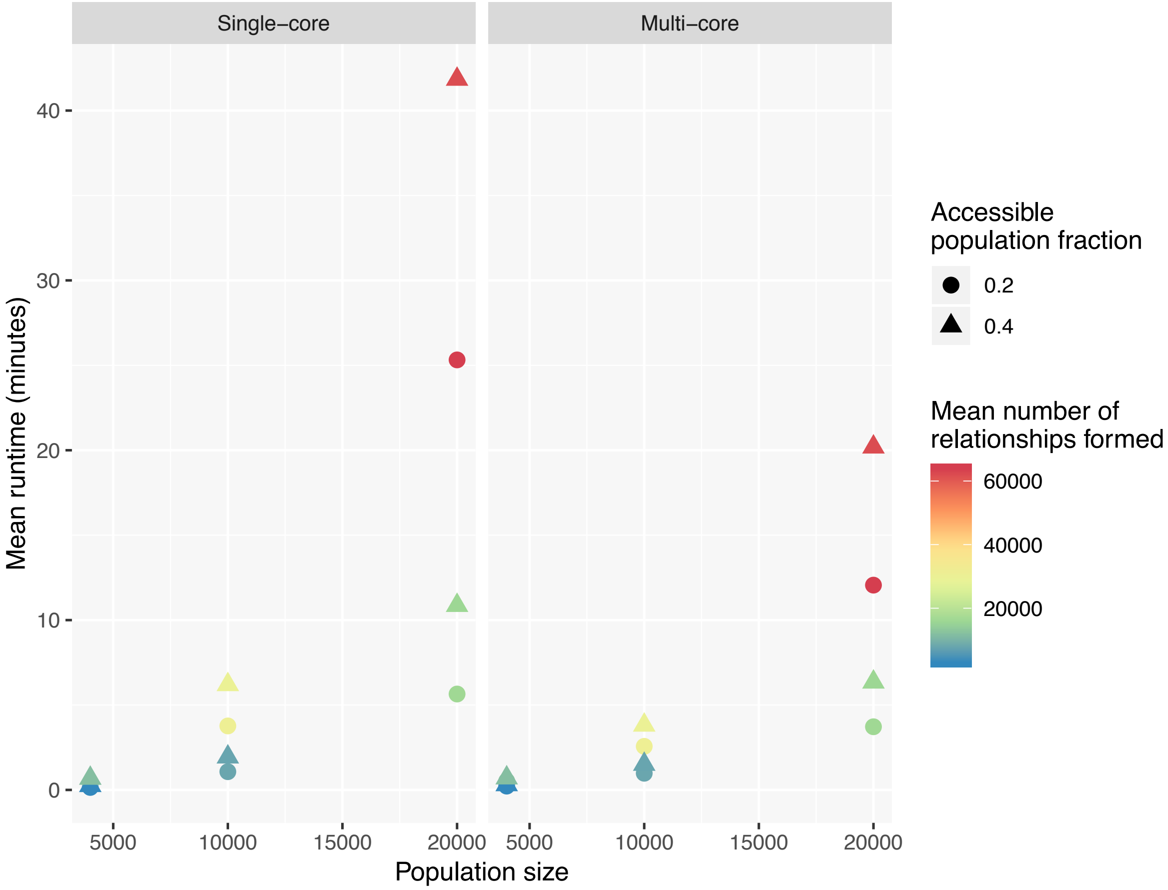

Figure 3. Mean runtime, in minutes, of simulation runs for model scenarios varying in population size, opposite sex population fraction accessible for relationship formation, mean number of relationships formed over the course of the simulation, and computing cores used for each of the runs.

The results of this simulation study are summarised in Figure 3. Runtimes increase faster than linearly with population size. This is to be expected, since the number of potential relationships that needs to be scheduled in a population of x men and y women is x*y. I.e. this number increases quadratically with population size. Reducing the number of potential partners by setting the `population.eyecap.fraction` parameter to a value smaller than 1, leads to an appreciable reduction in runtime, without affecting the rate at which relationships are formed by much. Running the simulation in parallel (all 6 cores of the machine's processor are used in the calculations) only leads to a (modest) speedup for scenarios of highly connected networks (many relationships are being formed) in large populations. Typically, simulation studies require several dozens or hundreds of simulation runs, and in that case it is more efficient to distribute single-core simulations in parallel over multiple cores than running multi-core simulations sequentially.

Model applications

The following section discusses two example simulations that were done using SimpactCyan. The first illustrates how SimpactCyan can be used to assess the impact of progressive changes to the ART eligibility criteria in Eswatini (formerly known as Swaziland). The second illustrates the use of SimpactCyan as a data-generating tool for assessing the performance of other modelling frameworks. All code and data files necessary to reproduce the examples are available at https://github.com/wdelva/SimpactCyanExamples.

The impact of Early Access to ART for All on HIV incidence

In the MaxART project[31], SimpactCyan is used to estimate the likely impact of Eswatini's shift towards ”Early Access to ART for All” (EAAA) on the incidence of HIV. HIV incidence is the rate at which HIV-uninfected people acquire the infection. Such infection events are scheduled each time a relationship is formed between an HIV-infected and an HIV-uninfected individual. The hazard for the event is given by

$$\text{hazard} = \exp(a + bV^{-c} + \text{other terms}),$$

where the other terms are not enabled by default, but allow for a hazard-lowering effect of multiple ongoing relationships (so-called coital dilution[32, 33]), as well as a hazard-increasing effect of adolescent age among women[34]. In this formula, a, b and c are model parameters; the V value represents the current HIV viral load of the person that is already infected.

The viral load model is based upon the notion that an infected person has a specific set-point viral load, V_{sp}, which corresponds to the viral load in the chronic stage of the infection. The three parameters `person.vsp.toacute.x`, `person.vsp.toaids.x` and `person.vsp.tofinalaids.x` determine the factors by which the HIV transmission hazard should be multiplied during the initial acute stage, as well as the early and late AIDS stages. The V value in this hazard expression can therefore be different from the V_{sp} value, depending on the time since infection. The non-linear form of this hazard function was inspired by equation (9) published by Hargrove et al.[35], while the default parameter values are based on a fit to model output generated by Fraser et al.[36].

At the time of HIV acquisition, time till HIV-related death is determined, based on a paper by Arnaout et al.[37]:

$$t_{\text{survival}} = \frac{C}{V_{\text{sp}}^{-k}} \times 10^{x}.$$

In this formula, C and k are parameters that can be configured by the user if desired; the x parameter (which defaults to zero) is person-specific, and its distribution can be configured to control the amount of variation in post-HIV infection survival times among people with the same set-point viral load.

The set-point viral load value allocated to a newly infected individual is partly determined by that of their infector, i.e. some heritability of set-point viral load is assumed[38]. This is done by using a two-dimensional distribution

$$\text{prob}(V_{\text{sp.infector}}, V_{\text{sp.infected}}),$$

of which the parameters can be chosen using the configuration values. This is subsequently used to obtain the conditional probability when fixing the initial V_{sp} value for a person that becomes infected due to the transmission event. By default, a symmetric, truncated bivariate normal distribution with mean 4, minimum 1, maximum 8, standard deviation 1 and correlation coefficient 0.33 is used to sample a $\log_{10}$ set-point viral load value for a newly infected person, conditional on the set-point viral load of the infector. To choose the initial set-point viral loads for ‘seed infections’, a marginal probability distribution is used, however.

ART initiation affects both the expected time till HIV-related death and the infectiousness of the person on ART. As soon as ART is started, the $\log_{10}$ viral load is assumed to drop by a user-defined fraction, and the updated current viral load is used to re-calculate t_{survival}. In the EAAA simulation study we assumed that upon ART initiation, the $\log_{10}$ viral load drops by 70%, effectively rendering the viral load “undetectable” for most ART clients. Via intervention events, most model parameters can be changed at arbitrary points in time during the simulation. However, person-specific parameter values (e.g. the probability of accepting ART if ART-eligible) and some event-times (e.g. time of non-HIV-related death) are determined at the time the individual is introduced into the population (at the start of the simulation or at birth). Hence, changing related parameters through an intervention event would only affect individuals born into the population after this intervention event, and not the extant population.

In this study, intervention events allowed us to assume that ART was gradually introduced around the year 2000, and that the CD4 cell count threshold for ART eligibility progressively shifted towards ever more inclusive criteria, alongside a decreasing lagtime between HIV infection and HIV diagnosis. These assumptions hold in both the ”Status Quo” scenario and the ”Early

Access to ART for All" (EAAA) scenario. In the EAAA scenario, however, an additional policy change is modelled: a policy of immediate access to ART for all people infected with HIV is adopted from October 2016. In the alternative scenario, the CD4 cell count threshold for ART eligibility stays at 500 cells/microliter from mid 2013 onwards.

The EAAA model was calibrated to demographic, epidemiological and programmatic data (which we refer to as target features) from Eswatini. Specifically, we used the UNAIDS annual national HIV prevalence estimates (1990-2017) and ART coverage estimates (2010-2017)[39], the estimated average population growth rate over the period 2000 to 2016[40], the gender- and age group-specific HIV prevalence and incidence estimates from the 2011-2012 SHIMS I study[41] and the UNAIDS 2017 estimate for the fraction of people aged 15 and above who were virally suppressed (less than 1000 viral copies per mL blood)[39]. Nineteen model parameters were calibrated to these data. Together these parameters determine the sexual behavioural and demographic dynamics of the model population, as well as temporal changes in the rate at which HIV-positive people in the model population were diagnosed with HIV infection, and the extent to which adolescent girls and young women are biologically more susceptible to HIV acquisition than older women and men. In doing so, these 19 parameters drive the model's features (i.e. summary statistics) that needed to be matched to the target features. Model calibration was achieved by applying the adaptive population Monte Carlo Approximate Bayesian Computation scheme describes by Lenormand et al.[42]. Iteratively sampling from the parameter space, starting from the prior distributions (wide-ranged uniform distributions) of the 19 model parameters, the method searches for areas in parameter space that produce model features close to the target features. After thirteen waves of simulations, totalling 29 500 model runs, the convergence criterion was reached and the calibration scheme produced a posterior distribution for the 19 parameters, by way of the 250 best fitting models. Here, "model" means a unique parameter combination producing model features similar to the target features. For each of the 250 models that jointly comprised the parameter posterior distribution, we calculated the root means squared relative error between model features and target features, as a summary measure of goodness-of-fit. The 3 models that fit the data best, which one can think of as the estimated mode of the posterior, were used in the forward projecting step of the analysis. In Supplementary Table S1, we provide the complete list of the 66 summary statistics (target features) that were used to calibrate the model, as well as the corresponding model features, obtained by averaging over the 10 model runs.

In this second part of the analysis, we simulated two scenarios for the expansion of ART in Eswatini, and for each of the two scenarios we ran each of the 3 models 10 times by keeping the model parameter fixed and only changing the seed of the random number generator. In the EAAA scenario, we simply used the 3 best-fitting models and ran them until 2032. In the counterfactual scenario, the same model parameters were used with the exception that the CD4 threshold for ART eligibility remained capped at 500 CD4 cells per microliter from mid 2013 onwards. In all of the subplots of Figure 4, the output from the 3 models is grouped by colour: shades of red for the EAAA scenario and shades of blue for the counterfactual scenario. For each model, the output of the 10 individual model runs is shown in thin dashed lines, and a solid line shows the average model trend. A darker, thick line represents the average across the 30 (3 times 10) runs for each scenario. Grey boxes represent the ranges around the UNAIDS estimates within which the actual numbers lie, based on the best available information[39]. The black dot in Figure 4b indicates the 2017 UNAIDS estimate for HIV incidence among 15-49 year-old adults in Eswatini[43] (not used for model calibration).

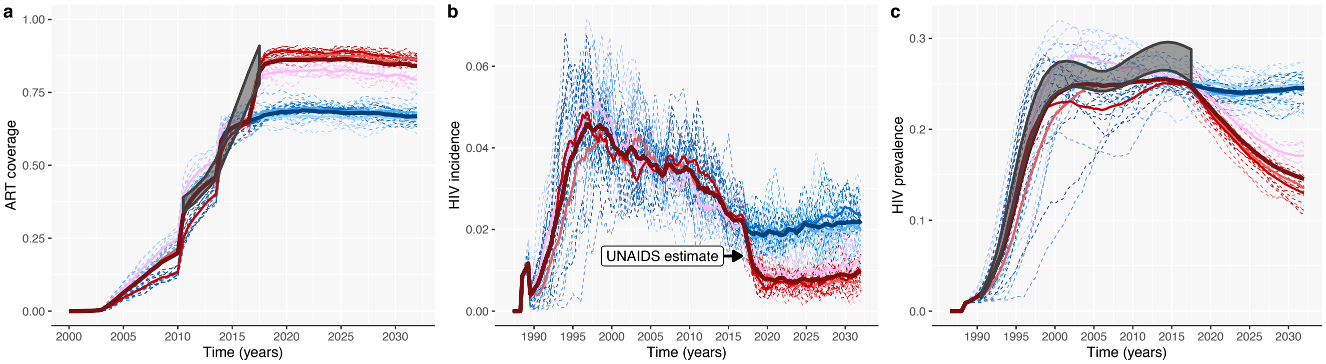

Figure 4. Programmatic and epidemic projections under a "Status Quo" (blue) and "Early Access to ART for All" (red) scenario for the roll-out of a nation-wide ART programme. (**a**) ART coverage: Fraction of the adult HIV-positive population (≥ 15 years old) on ART. (**b**) HIV incidence rate among 15 to 49 year-old people. (**c**) HIV prevalence among 15 to 49 year-old people.

The impact of the policy shift to EAAA was estimated by the relative reduction in HIV incidence (1 minus the ratio of the incidence rates under the two scenarios). Under the counterfactual scenario, HIV incidence was projected to drop by 12%

by the end of 2019, from its base level of 2.2 / 100 PY in October 2016. However, under the factual scenario, the models, on average, estimated that the incidence will decrease by 64% over that same period, to 0.8 / 100 PY. The impact of EAAA, as measured by the incidence rate ratio for the 2 scenarios is projected to remain stable over the next decade. The variation between the output of these three models, as shown in Figure 4b, provides a sense of the uncertainty around our best guess of the future impact of EAAA on HIV incidence. It should, however, not be interpreted as an estimate of the credibility or confidence interval, because only the mode of the posterior was used in the impact estimation.

SimpactCyan as a data-generating benchmarking tool

The second use case illustrates how SimpactCyan can be used as a data-generating tool for benchmarking the performance of other modelling frameworks. Phylogenetic models have been used to infer properties of epidemics from reconstructed phylogenetic trees, including time-trends in HIV incidence rates[44,45] and the age-mixing pattern in HIV transmission clusters[46]. Yet, as the truth is typically unknown, it is difficult to assess the validity of these novel modelling frameworks, or document their sensitivity to breaches in the models' assumptions. For instance, some phylogenetic inference methods require that HIV sequence data are available for the majority of HIV-positive people[47]. However, in many settings, this assumptions is not met. In Figure 5 we illustrate the basic idea of SimpactCyan as a data-generating tool for benchmarking.

First, we simulated an HIV epidemic similar to the epidemics in the EAAA analysis explained in the previous model application. Figure 5a shows the cumulative HIV transmission network that originated from one of the 20 seed infectors (man 858). This network links all individuals who got infected with HIV by man 858 or one of his descendants by the end of the simulation. Next, we converted the transmission network into a phylogenetic tree by assuming that HIV transmission events correspond to branching points in the phylogeny. While this assumption is obviously overly simplistic, it is not an uncommon assumption in phylodynamic modelling studies[48–51]. More importantly, it does not invalidate the didactic purpose of this model application, which is to show how SimpactCyan can be used as a data-generating tool for benchmarking the performance of other modelling frameworks in simulation settings where the user has control over the extent to which the assumptions made by the phylodynamic inference method are in line with the processes that generated the simulated data. Using the Seq-Gen program[52], we simulated the viral evolution along this phylogeny, assuming a generalised time-reversible substitution model[53] with a gamma-invariable mixture model (GTR+γ+I) for rate heterogeneity among sites. In this way, we generated synthetic HIV sequence data. Specifically, for the root sequence we chose a consensus sequence of the pol region of HIV-1 subtype C, isolated in 1989 from South Africa and retrieved from the HIV Sequence Database of the Los Alamos National Library (LANL) (access number HIV.1.C.ZA.PolCDS1989)[54]. The program jModelTest version 2.1.3[55] was used for selecting the best fitting evolutionary model to explain the viral diversity in a dataset of HIV-1 subtype C sequences from 386 South African patients, also extracted from the LANL HIV Sequence Database. Using hierarchical likelihood ratio tests, the GTR+γ+I model was ranked as the best fitting evolutionary model for these viral sequence data, and the model's parameter values as estimated by jModelTest were used in the forward viral evolution simulation. Specifically, the relative frequencies of adenine (A), cytosine (C), guanine (G), and thymine (T) were estimated at 0.3906, 0.1752, 0.2201, and 0.2142 respectively. Further, the inferred value for the shape parameter for the γ rate heterogeneity was 0.625, and the heterogeneity in transition rate across sites was discretized into 4 discrete γ rate categories. The 6 substitution rate parameters of the GTR model were estimated at 1.9803, 9.4404, 0.9423, 0.8770, 11.6367, and 1.0000, with an assumed evolutionary rate of 0.00475 substitutions per site per year (branch length scaling factor)[56,57]. The fraction of invariant sites was estimated at 0.213. Next, we fed these synthetic sequences into the phangorn[58] and treedater R packages[59] to reconstruct the time-resolved phylogenetic tree (Figure 5b), by fitting the GTR+γ+I model with a likelihood-based approach and root-to-trip regression. Lastly, we summarised the timing of the internal nodes in the reconstructed time-resolved tree by a vector of the number of internal nodes (i.e. branching points) per calendar year time interval.

In Supplementary Table S2, we show the model parameters of the generalised time-reversible substitution model with a gamma-invariable mixture model (GTR+γ+I) for rate heterogeneity among sites, fitted to the empirical sequence data (sample size = 386), and fitted to the simulated sequences (sample size = 2896). To allow for a more direct comparison, we extracted from the complete synthetic sequence dataset a matching sample of simulated sequences with sampling dates as close as possible to those of the empirical sequence data. In Supplementary Figure S1, we also show the empirical phylogenetic tree, the matching sample simulated tree, and a violin plot of the density of patristic distances of the respective trees. Lastly, we report in Supplementary Table S3 nine topological properties of the respective trees. Taken together, this additional information shows strong agreement between empirical and simulated data.

If all people ever infected are included in the sequence dataset and the same molecular evolution model is used to generate the sequence data and to reconstruct the phylogenetic tree, the timing of the internal nodes in the reconstructed tree should correspond with the timing of the simulated HIV transmission events. Indeed, in a perfect scenario of internal consistency and complete data, the distribution of internal nodes in the reconstructed phylogenetic tree and simulated HIV transmission events matched nearly perfected (red and dark blue lines in Figure 5c). While an exhaustive sensitivity analysis of how phylodynamic

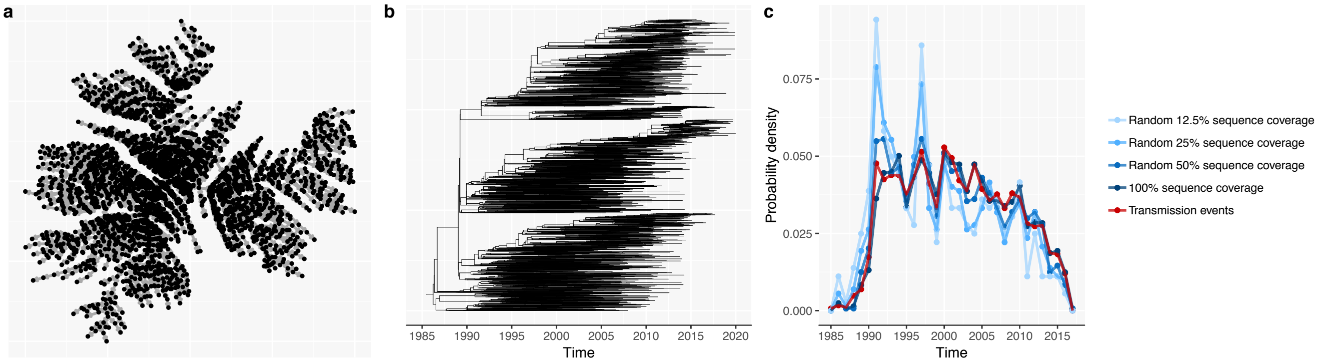

Figure 5. (**a**) The cumulative HIV transmission network, linking all individuals whose HIV infections originated from one seed infector (man 858). (**b**) The time-resolved phylogenetic tree, reconstructed from synthetic HIV sequence data, generated by simulating the molecular evolution of HIV viral strains across the HIV transmission network. (**c**) The probability density of internal nodes in the reconstructed phylogenetic tree(shades of blue) and simulated HIV transmission events (red). The correlation between the timing of the internal nodes and HIV transmission events becomes weaker as the HIV sequence coverage decreases.

inference could be affected by missing data and assumptions that are not consistent with the data-generating processes is beyond the scope of this paper, we simulated three additional scenarios, to illustrate how reduced sequences coverage (50%, 25% and 12.5%) could add noise and bias to the phylodynamic inference. Coverage here is defined as the fraction of people in the cumulative HIV transmission network for whom a consensus sequence is included in the HIV sequence database. In all three of these imperfect scenarios, we still used the appropriate molecular evolution model to reconstruct the phylogenetic tree. As sequence coverage decreases, the timing of internal nodes becomes a less accurate proxy for the timing of transmission events, and hence, a less reliable source for inferring time trends in HIV incidence.

Future directions

Ongoing developments of SimpactCyan include the addition of events for the transmission and treatment of other sexually transmitted infections such as Herpes Simplex Virus 2 (HSV-2) and Hepatitis C Virus (HCV), as well as additional events for parenteral and mother-to-child transmission of HIV and co-infections, to allow studies of HIV transmission in injecting drug users (IDU) and children. We also plan to extend the software by enabling explicit modelling of relationships between sexual risk behaviour and health seeking behaviour. This is in response to recent evidence to suggest that high sexual risk behaviour is associated with a lower likelihood to be aware of one's HIV infection, and a lower likelihood of being virally suppressed among people who know they are HIV positive[60].

Conceived as a flexible open-source, open access tool, rather than a proprietary asset, SimpactCyan's extensions and applications should not solely come from its original developers. Instead, we want to position this simulator as a vehicle for open science in HIV epidemiology. Therefore, others are encouraged to use it for the development of their own IBMs, as the starting point for their own simulation engine, as a data-generating and/or benchmarking tool in methodological research, or for educational purposes.

Acknowledgements

The authors thank Jonathan Dushoff and Roxanne Beauclair for valuable comments on earlier drafts of this manuscript. The computational resources (Stevin Supercomputer Infrastructure) and services used for the calibration of the EAAA model were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

Funding sources

This work was supported by grants G091210N, G0B4314N and W002514N from the Research Foundation – Flanders (FWO), grant ZEIN2010PR375 from the Flemish Interuniversity Council (VLIR), grant 100014 from NRF-TWAS, and a grant from the Dutch Postcode Lottery.

Author contributions

JL wrote all the source code of the core engine and the R and Python interfaces presented in this paper, and wrote the first draft of the manuscript. DMH drafted the introduction section and Figure 2. EK contributed to Figure 2 and provided editorial assistance. DN wrote R code for the second example application. NH contributed to the software design and provided editorial assistance. WD contributed to the software design, performed the runtime analysis, wrote R code for the example applications, and wrote the ”Model applications” and ”Future directions” sections of the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Railsback, S. F. & Grimm, V. *Agent-based and individual-based modeling: a practical introduction* (Princeton university press, 2011).

2. Willem, L., Verelst, F., Bilcke, J., Hens, N. & Beutels, P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). *BMC Infect. Dis.* **17**, 612 (2017). DOI 10.1186/s12879-017-2699-8.

3. Kravari, K. & Bassiliades, N. A survey of agent platforms. *J. Artif. Soc. Soc. Simul.* **18**, 11 (2015).

4. Delva, W., Hazelbag, C. M. & Dushoff, J. Simulation-based calibration of individual-based models to summarised data in hiv epidemiology: a systematic literature review and new mice-based method. In *22nd International AIDS Conference (AIDS 2018)* (Amsterdam, The Netherlands, 2018).

5. Liesenborgs, J. Simpact cyan. https://simpactcyan.readthedocs.io/en/latest/index.html (2017). Accessed: 2019-04-10.

6. Liesenborgs, J., Hendrickx, D. M., Van Reeth, F., Hens, N. & Delva, W. Simpactcyan: a simulation engine for individual-based models in hiv epidemiology. In *Epidemics6 – 6th International Conference on Infectious Disease Dynamics* (Sitges, Spain, 2017).

7. Hendrickx, D. M. *et al.* Comparison of two individual-based models for simulating hiv epidemics in a population with hsv-2 using as case study yaoundé-cameroon, 1989-1998. *bioRxiv* (2019). DOI 10.1101/637389. https://www.biorxiv.org/content/early/2019/05/17/637389.full.pdf.

8. Sahneh, F. D., Vajdi, A., Shakeri, H., Fan, F. & Scoglio, C. GEMFsim: A stochastic simulator for the generalized epidemic modeling framework. *J. Comput. Sci.* (2017). DOI 10.1016/j.jocs.2017.08.014.

9. Moshiri, N., Ragonnet-Cronin, M., Wertheim, J. O. & Mirarab, S. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics* (2018). DOI 10.1093/bioinformatics/bty921.

10. Jenness, S. M., Goodreau, S. M. & Morris, M. EpiModel : An R Package for Mathematical Modeling of Infectious Disease over Networks . *J. Stat. Softw.* (2018). DOI 10.18637/jss.v084.i08.

11. Bakker, R. *et al.* Stdsim: A microsimulation model for decision support in the control of hiv and other stds. *Sex. Transm. Dis.* **27**, 652 (2000).

12. Thiele, J. C. R marries netlogo: Introduction to the rnetlogo package. *J. Stat. Softw.* **58** (2014). DOI 10.18637/jss.v058.i02.

13. Reniers, G., Armbruster, B. & Lucas, A. Sexual networks, partnership mixing, and the female-to-male ratio of HIV infections in generalized epidemics: An agent-based simulation study. *Demogr. Res.* **33**, 425–450 (2015). DOI 10.4054/DemRes.2015.33.15.

14. Bershteyn, A. *et al.* Implementation and applications of emod, an individual-based multi-disease modeling platform. *Pathog. Dis.* **76** (2018). DOI 10.1093/femspd/fty059.

15. McCormick, A. W. *et al.* Development, calibration and performance of an HIV transmission model incorporating natural history and behavioral patterns: Application in South Africa. *PLoS ONE* **9** (2014). DOI 10.1371/journal.pone.0098272.

16. Johnson, L. F., Kubjane, M. & Moolla, H. Microcosm: a model of social and structural drivers of hiv and interventions to reduce hiv incidence in high-risk populations in south africa. https://www.biorxiv.org/content/10.1101/310763v1 (2018). Accessed: 2019-04-10.

17. Gopalappa, C., Farnham, P. G., Chen, Y.-H. & Sansom, S. L. Progression and transmission of hiv/aids (path 2.0): A new, agent-based model to estimate hiv transmissions in the united states. *Med. Decis. Mak.* **37**, 224–233 (2017). DOI 10.1177/0272989X16668509. PMID: 27646567, https://doi.org/10.1177/0272989X16668509.

18. Marshall Labs. Treatment of infectious transmissions through agent-based network. https://titan-documentation.readthedocs.io/en/latest/index.html (2017). Accessed: 2019-04-10.

19. Tolentino, S. L., Meng, F. & Delva, W. A simulation-based method for efficient resource allocation of combination hiv prevention. In *Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques (Cannes, France, 5-7 March 2013)*, 31–40 (ICST, Brussels, Belgium, 2013).

20. Meng, F., Hummeling, R., Tolentino, S. L., Hens, N. & Delva, W. Modelling the impact of alternative treatment strategies on hiv prevalence in south africa: a simulation study. In *6th South African AIDS Conference* (Durban, South Africa, 2013).

21. Delva, W. *et al.* Age mixing and sustained hiv epidemics: not the size but the variation of age gaps counts. In *Epidemics4 – 4th International Conference on Infectious Disease Dynamics* (Amsterdam, The Netherlands, 2013).

22. Tolentino, S. L. *Effective and efficient algorithms for simulating sexually transmitted diseases*. Ph.D. thesis, University of Iowa (2014).

23. Meng, F., Liesenborgs, J., Delva, W., Van Reeth, F. & Hens, N. Simpact cyan: accelerating agent-based, continuous time simulation of hiv transmission on vsc. In *VSC Users Day 2014* (Brussels, Belgium, 2014).

24. Anderson, D. F. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The J. Chem. Phys.* **127** (2007). DOI http://dx.doi.org/10.1063/1.2799998.

25. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**, 403 – 434 (1976). DOI http://dx.doi.org/10.1016/0021-9991(76)90041-3.

26. Bartlett, M. S. Stochastic Processes or the Statistics of Change. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **2**, 44–64 (1953). DOI 10.2307/2985327.

27. Doob, J. L. Topics in the theory of markoff chains. *Transactions Am. Math. Soc.* **52**, 37–64 (1942). DOI 10.1090/S0002-9947-1942-0006633-7.

28. Gibson, M. A. & Bruck, J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The J. Phys. Chem. A* **104**, 1876–1889 (2000). DOI 10.1021/jp993732q. http://dx.doi.org/10.1021/jp993732q.

29. Hollingsworth, T. D., Anderson, R. M. & Fraser, C. HIV-1 Transmission, by Stage of Infection. *The J. Infect. Dis.* **198**, 687–693 (2008). DOI 10.1086/590501.

30. Grimes, R. M., Hallmark, C. J., Watkins, K. L., Agarwal, S. & McNeese, M. L. Re-engagement in hiv care: A clinical and public health priority. *J AIDS Clin Res* **7**, 1–7 (2016). DOI 10.4172/2155-6113.1000543.

31. Walsh, F. J. *et al.* Impact of early initiation versus national standard of care of antiretroviral therapy in Swaziland ' s public sector health system : study protocol for a stepped-wedge randomized trial. *Trials* **18**, 1–10 (2017). DOI 10.1186/s13063-017-2128-8.

32. Delva, W. *et al.* Coital frequency and condom use in monogamous and concurrent sexual relationships in cape town, south africa. *J. Int. AIDS Soc.* **16** (2013).

33. Sawers, L., Isaac, A. G. & Stillwaggon, E. Hiv and concurrent sexual partnerships: modelling the role of coital dilution. *J. Int. AIDS Soc.* **14**, 44 (2011).

34. Yi, T. J., Shannon, B., Prodger, J., McKinnon, L. & Kaul, R. Genital immunology and HIV susceptibility in young women. *Am. J. Reproductive Immunol.* **69**, 74–79 (2013). DOI 10.1111/aji.12035.

35. Hargrove, J., Eastwood, H., Mahiane, G. & van Schalkwyk, C. How should we best estimate the mean recency duration for the bed method? *PLOS ONE* **7**, 1–12 (2012). DOI 10.1371/journal.pone.0049661.

36. Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F. & Hanage, W. P. Variation in hiv-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. *Proc. Natl. Acad. Sci.* **104**, 17441–17446 (2007). DOI 10.1073/pnas.0708559104. http://www.pnas.org/content/104/44/17441.full.pdf.

37. Arnaout, R. A. *et al.* A simple relationship between viral load and survival time in hiv-1 infection. *Proc. Natl. Acad. Sci.* **96**, 11549–11553 (1999). DOI 10.1073/pnas.96.20.11549. http://www.pnas.org/content/96/20/11549.full.pdf.

38. Fraser, C. *et al.* Virulence and pathogenesis of hiv-1 infection: an evolutionary perspective. *Science* **343**, 1243727 (2014).

39. UNAIDS. AIDSinfo. http://aidsinfo.unaids.org/ (2019). Accessed: 2019-04-20.

40. World Bank. https://data.worldbank.org/indicator/SP.POP.GROW?locations=SZ (2019). Accessed: 2019-04-24.

41. Justman, J. *et al.* Swaziland hiv incidence measurement survey (shims): a prospective national cohort study. *The lancet HIV* **4**, e83–e92 (2017).

42. Lenormand, M., Jabot, F. & Deffuant, G. Adaptive approximate Bayesian computation for complex models. *Comput. Stat.* **28**, 2777–2796 (2013). DOI 10.1007/s00180-013-0428-3.

43. UNAIDS. https://data.worldbank.org/indicator/SH.HIV.INCD.ZS?locations=SZ (2018). Accessed: 2019-04-26.

44. Rasmussen, D. A., Volz, E. M. & Koelle, K. Phylodynamic Inference for Structured Epidemiological Models. *PLoS Comput. Biol.* **10** (2014). DOI 10.1371/journal.pcbi.1003570.

45. Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A. & Leigh Brown, A. J. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Medicine* **5**, 0392–0402 (2008). DOI 10.1371/journal.pmed.0050050.

46. de Oliveira, T. *et al.* Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *The Lancet HIV* **3018**, 1–10 (2016). DOI 10.1016/S2352-3018(16)30186-2.

47. Novitsky, V., Moyo, S., Lei, Q., DeGruttola, V. & Essex, M. Impact of Sampling Density on the Extent of HIV Clustering. *AIDS Res. Hum. Retroviruses* **30**, 1226–1235 (2014). DOI 10.1089/aid.2014.0173.

48. Leventhal, G. E., Günthard, H. F., Bonhoeffer, S. & Stadler, T. Using an epidemiological model for phylogenetic inference reveals density dependence in hiv transmission. *Mol. Biol. Evol.* (2014). DOI 10.1093/molbev/mst172.

49. Ratmann, O. *et al.* Phylogenetic tools for generalized HIV-1 epidemics: Findings from the PANGEA-HIV methods comparison. *Mol. Biol. Evol.* (2017). DOI 10.1093/molbev/msw217.

50. Delva, W., Leventhal, G. E. & Helleringer, S. Connecting the dots: network data and models in hiv epidemiology. *AIDS* **30**, 2009–2020 (2016).

51. Villandre, L. *et al.* Assessment of overlap of phylogenetic transmission clusters and communities in simple sexual contact networks: Applications to HIV-1. *PLoS ONE* (2016). DOI 10.1371/journal.pone.0148459.

52. Rambaut, A. & Grassly, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. applications biosciences : CABIOS* **13**, 235–8 (1997). DOI 10.1093/BIOINFORMATICS/13.3.235.

53. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. on mathematics life sciences. Vol. 17* 57–86 (1986). DOI citeulike-article-id:4801403.

54. Los Alamos National Library. HIV Sequence database. https://www.hiv.lanl.gov/components/sequence/HIV/search/search.html (2018). Accessed: 2018-07-12.

55. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jmodeltest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 EP – (2012).

56. Patiño-Galindo, J. Á. & González-Candelas, F. The substitution rate of HIV-1 subtypes: a genomic approach. *Virus Evol.* (2017). DOI 10.1093/ve/vex029.

57. Alizon, S. & Fraser, C. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology* (2013). DOI 10.1186/1742-4690-10-49.

58. Schliep, K. P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011). DOI 10.1093/bioinformatics/btq706.

59. Volz, E. M. & Frost, S. D. W. Scalable relaxed clock phylogenetic dating. *Virus Evol.* **3** (2017). DOI 10.1093/ve/vex025.

60. Huerga, H. *et al.* Higher risk sexual behaviour is associated with unawareness of hiv-positivity and lack of viral suppression –implications for treatment as prevention. *Sci. Reports* **7**, 16117 (2017). DOI 10.1038/s41598-017-16382-6.