Why the proportion of transmission during early-stage HIV infection does not predict the long-term impact of treatment on HIV incidence

Jeffrey W. Eaton1 and Timothy B. Hallett

Department of Infectious Disease Epidemiology, Imperial College London, St Mary’s Hospital, London W2 1PG, United Kingdom

Edited by Alan S. Perelson, Los Alamos National Laboratory, Los Alamos, NM, and accepted by the Editorial Board September 5, 2014 (received for review December 23, 2013)

Antiretroviral therapy (ART) reduces the infectiousness of HIV-infected persons, but only after testing, linkage to care, and successful viral suppression. Thus, a large proportion of HIV transmission during a period of high infectiousness in the first few months after infection (“early transmission”) is perceived as a threat to the impact of HIV “treatment-as-prevention” strategies. We created a mathematical model of a heterosexual HIV epidemic to investigate how the proportion of early transmission affects the impact of ART on reducing HIV incidence. The model includes stages of HIV infection, flexible sexual mixing, and changes in risk behavior over the epidemic. The model was calibrated to HIV prevalence data from South Africa using a Bayesian framework. Immediately after ART was introduced, more early transmission was associated with a smaller reduction in HIV incidence rate—consistent with the concern that a large amount of early transmission reduces the impact of treatment on incidence. However, the proportion of early transmission was not strongly related to the long-term reduction in incidence. This was because more early transmission resulted in a shorter generation time, in which case lower values for the basic reproductive number ($R_0$) are consistent with observed epidemic growth, and $R_0$ was negatively correlated with long-term intervention impact. The fraction of early transmission depends on biological factors, behavioral patterns, and epidemic stage and alone does not predict long-term intervention impacts. However, early transmission may be an important determinant in the outcome of short-term trials and evaluation of programs.

HIV incidence | antiretroviral therapy | early infection | HIV prevention intervention | mathematical model

Recent studies have confirmed that effective antiretroviral therapy (ART) reduces the transmission of HIV among stable heterosexual couples (1–3). This finding has generated interest in understanding the population-level impact of HIV treatment on reducing the rate of new HIV infections in generalized epidemic settings (4). Research, including mathematical modeling (5–10), implementation research (11), and major randomized controlled trials (12–14), are focused on how ART provision might be expanded strategically to maximize its public health benefits (15, 16).

One concern is that if a large fraction of HIV transmission occurs shortly after a person becomes infected, before the person can be diagnosed and initiated on ART, this will limit the potential impact of HIV treatment on reducing HIV incidence (9, 17, 18). Data suggest that persons are more infectious during a short period of “early infection” after becoming infected with HIV (19–22), although there is debate about the extent, duration, and determinants of elevated infectiousness (18, 23). The amount of transmission that occurs also will depend on patterns of sexual behavior and sexual networks (17, 24–27). There have been estimates for the contribution of early infection to transmission from mathematical models (7, 17, 21, 24–26) and phylogenetic analyses (28–31), but these vary widely, from 5% to above 50% (23).

In this study, we use a mathematical model to quantify how the proportion of transmission that comes from persons who have been infected recently affects the impact of treatment scale-up on HIV incidence. The model is calibrated to longitudinal HIV prevalence data from South Africa using a Bayesian framework. Thus, the model accounts for not only the only the epidemic growth rate highlighted in previous research (5, 9, 18), but also the heterogeneity and sexual behavior change to explain the peak and decline in HIV incidence observed in sub-Saharan African HIV epidemics (32, 33).

The model calibration allows uncertainty about factors that determine the amount of early transmission, including the relative infectiousness during early infection, heterogeneity in propensity for sexual risk behavior, assortativity in sexual partner selection, reduction in risk propensity over the life course, and population-wide reductions in risk behavior in response to the epidemic (32, 33). This results in multiple combinations of parameter values that are consistent with the observed epidemic and variation in the amount of early transmission. We simulated the impact of a treatment intervention and report how the proportion of early transmission correlates with the reduction in HIV incidence from the intervention over the short- and long-term.

Results

Model Calibration and Transmission During Epidemic Stage. The mathematical model was calibrated to the time series of HIV prevalence among pregnant women in South Africa from 1990 through 2008 (34) and HIV prevalence among adult men and women aged 15–49 y in three national household surveys in 2002, 2005, and 2008 (Fig. 1A). The HIV incidence rate grows rapidly in the early 1990s, peaking in 1997 and declining by around 37% (95% credible interval 30–43) over the next decade.

Significance

Antiretroviral treatment (ART), which prevents both morbidity and HIV transmission for persons infected with HIV, is now thought to be central to strategies for controlling the spread of HIV. One major concern has been that if a large proportion of HIV transmission occurs early in infection, before persons can be diagnosed and treated, the impact of treatment on reducing new infections will be less. We use a mathematical model to quantify how the proportion of early transmission will affect the impact of intervention strategies, and explain underlying epidemiological mechanisms for this. Model simulations suggest that—counter to expectations—the proportion of early transmission at the start of an ART intervention does not predict the long-term intervention impact.

Author contributions: J.W.E. and T.B.H. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

This article is a PNAS Direct Submission. A.S.P. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

1To whom correspondence should be addressed. Email: jeffrey.eaton@imperial.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1323007111/-/DCSupplemental.
(Fig. 1B). The decline in incidence is the result of two factors: infection saturation in the higher-risk groups and a reduction in the unprotected sexual contact rate over time. It is estimated that the contact rate declines by 28% (95% CI, 11–43; parameter posterior distributions in Fig. S1).

The fraction of transmission occurring from individuals in each stage of infection evolves over the course of the epidemic (Fig. 1C). During the early growth phase of the epidemic, most infected persons have been infected recently, and so most new infections result from highly infectious persons in early HIV infection (early transmission). As the epidemic matures, the contribution of early transmission declines whereas the contribution of advanced stages of infection grows.

In the absence of ART scale-up, early transmission was estimated to account for 17% (95% CI, 12–24) of all transmission in 2010 (Fig. 2A). Much of the variation in early transmission was explained by the relative infectiousness during early infection compared to chronic infection ($R^2 = 56\%$; Fig. 2B). Adjusting for the influence of infectiousness during early infection, higher rates of moving from higher- to lower-risk groups resulted in more early transmission because faster movement to low-risk groups results in more transmission occurring early after infection. These two parameters together explained 89% of the variation in early transmission in 2010 (Fig. S2A).

**Early Transmission and the Impact of Treatment as Prevention.** We explored how the proportion of early transmission in 2010 affected the impact of ART on reducing HIV incidence. We simulated an intervention in which infected persons are eligible for treatment when their CD4 cell count falls below 350 cells/$\mu$L. Eligible persons initiated ART at a rate of 0.23 per year, such that 80% will initiate treatment before death. The intervention was introduced into an ART-naive epidemic in 2010, and the intervention was simulated for 30 y. We calculated the percentage reduction in incidence rate with reference to the model projection without ART.

**$R_0$ Early Transmission, Epidemic Growth, and Intervention Impact.** We calculated the basic reproductive number ($R_0$; the average number of new infections that one infected individual would create in a wholly susceptible population) over the course of the epidemic. The posterior mean for $R_0$ at the start of the epidemic was 4.9 (95% CI, 3.8–6.2) (Fig. 4A). After the estimated reduction in contact rate, $R_0$ during the intervention period was 3.5 (2.4–5.0).

There was a strong negative relationship between the initial $R_0$ and early transmission in 2010 (Fig. 4B) (9, 18). Eighty-six percent of the variation in the proportion of early transmission is explained by the reciprocal of $R_0$ ($1/R_0$; red line in Fig. 4B). This was because when there is more early transmission, on average onward transmissions occur sooner after an individual becomes infected—that is, the “generation time” is shorter. With shorter generation times, the observed rapid epidemic growth rate in HIV prevalence may be achieved with lower values of $R_0$ (18, 35).

The variation in $R_0$ at the start of the epidemic is explained predominantly by two parameters—the relative infectiousness during early infection and the rate of risk group movement—which jointly accounted for 92% of the variation (Fig. S2B). The reduction in the contact rate also is an important determinant of $R_0$ during the intervention period, and together those three parameters accounted for 95% of the variation in $R_0$ during the intervention period (Fig. S2C).

The lack of association between the proportion of early transmission and the long-term reduction in HIV incidence due to ART therefore may be understood with reference to $R_0$. If a larger fraction of transmission occurs during early transmission—before individuals receive ART, limiting the impact of ART—then it also will be true that $R_0$ is lower, which means any intervention may have a greater long-term impact on incidence (Fig. 4C) (36). These two opposing influences appear to counterbalance each other, resulting in the lack of association between early transmission and
the long-term impact of treatment interventions on incidence (Fig. 3).

**The Influence of Elevated Infectiousness During Early Infection.** The extent to which individuals in early infection are more infectious is highly uncertain. However, we observe no relationship between the parameter determining the increased infectiousness during early infection and the long-term impact on HIV incidence of treatment interventions. (19) In a secondary analysis, we calibrated model parameters conditional on fixed values of the relative early infectiousness, assuming that persons in early infection were (i) 26 times more infectious (21), (ii) 9.2 times more infectious (37); and (iii) had the same average infectiousness (as assumed by ref. 5) compared with persons with CD4 >350 cells/μL (calibration, Fig. S4; posterior distributions, Fig. S5). These different fixed levels for relative early infectiousness did not affect the impact of the ART intervention on reducing incidence after 30 y (Fig. S5B). The reason is the same as above: higher relative infectiousness resulted in more early transmission in the model and therefore lower R0 values after calibration to observed epidemic data (Table 1).

**Discussion**

The share of HIV transmission during early infection and the implications for epidemic control have attracted substantial empirical and theoretical attention. Examining these issues within a flexible mathematical model of a heterosexual HIV transmission and calibrating this to the growth, peak, and decline of a well-documented HIV epidemic have illuminated several important concepts. First, the fraction of transmission during early infection should be considered a quantity that emerges from a complex interaction of behavioral and biological factors rather than a fixed epidemiological characteristic of a given epidemic setting (38). This quantity changes with epidemic stage, epidemiological context, and the stage of the HIV treatment scale-up. Empirical measurement and comparisons of the contribution of early transmission over time and across epidemic settings must be interpreted with these factors in mind.

Second, over the short term, the model predicted that more early transmission would be associated with a smaller reduction in incidence rate following the scale-up of an intervention. Thus, the proportion of new infections arising from early infection may be useful in interpreting the results of community-based trials of treatment, which will measure impact on cumulative incidence over a 2–3-y period.

Third, the proportion of transmission occurring during early infection is surprisingly not predictive of the long-term impact of treatment interventions in model simulations. This is because, through calibration to observed data, epidemiologic parameters that created a larger amount of transmission in early infection also generated lower values of R0. Lower values of R0 mean that the same intervention may have a larger impact. This counterbalances the effect that more transmission in early infection interventions weakens interventions that block transmission during chronic infection.

This analysis has focused on whether the proportion of early transmission during the current endemic stage is predictive of the short- and long-term impact of a given intervention strategy in a generalized, predominantly heterosexually transmitted HIV epidemic setting. It did not evaluate what level of intervention would be required to eliminate HIV using treatment, and so our results are not directly comparable to studies that have evaluated thresholds for HIV elimination using ART [e.g., Granich et al. (5), Powers et al. (17), Hontelez et al. (8), Kretzschmar et al. (9)]. The results also may not apply to different epidemiological contexts; for example, concentrated epidemic settings in which transmission occurs primarily among men who have sex with men or persons who inject drugs. In reality, HIV incidence trends during the ART era will depend on many other factors, including changes in risk behavior (39, 40). The abstract representation of sexual risk, mixing, behavior change, and future epidemiologic
changes may limit the usefulness of this model for projecting specific intervention impacts and allocating resources in specific settings compared with other models with detailed demographic, spatial, and risk group structures.

Finally, although it is unknown by how much transmission is elevated during early infection, this does not materially add to uncertainty about the impact of treatment on incidence. Uncertainties about other factors (especially sexual behavior and mixing) mean that the observed epidemic may be “explained” in many ways whatever the value for the increased level of infectiousness. It is this latitude that drives uncertainty in estimates. Therefore, further measurement of infectiousness during early-stage infection would not be a priority from the perspective of informing projections of the long-term impact of treatment.

Previous work illustrated how different patterns of heterogeneity in sexual mixing and assumptions about changes in individual risk behavior may give rise to very different endemic prevalence for the same average contact rates or biological assumptions about infectiousness (24, 25, 27, 41–43). This analysis considers similar ideas from a different angle; it treated the prevalence as known and explored consistent combinations of behavioral and biological parameters. Two behavioral parameters—the rate of transition from higher-to lower-risk groups and the population-level reduction in unprotected contact rate over time—were particularly important for simulating the observed prevalence trend in many different ways, as well as determining the intervention impact. There is evidence that reductions in risk have contributed to changes in HIV both in South Africa (8, 33) and elsewhere (32, 44, 45). The value of $R_0$ during the intervention period was much more predictive of the reduction in incidence than $R_0$ at the start of the epidemic (Fig. 4). This finding suggests that models that appeal solely to estimates of $R_0$ based on initial epidemic growth rate to estimate intervention impact or consider only uncertainty about biological determinants of transmission, without accounting for any changes in behavior required to explain the peak and stabilization of the epidemic, may not provide robust insights about the likely impact of HIV prevention programs today, including ART.

On the other hand, the influence of calibrating the model to the historical epidemic data for identifying plausible combinations of parameter values that were used for projecting the long-term consequences of the intervention suggests that predictions based on models that simulate a plausible current epidemic level but do not simulate a credible epidemic growth (24, 25, 46, 47) should be treated with caution. It is not feasible to measure $R_0$ directly in an endemic HIV setting, but continued surveillance of all the factors that determine $R_0$—biological and behavioral determinants of transmission—will be critical for understanding and projecting the implications of control strategies, rather than relying on a single metric, such as the proportion of early transmission.

**Conclusion**

Modeling illustrated that a large fraction of transmission during early infection resulted in a smaller impact of treatment on HIV incidence in the first few years after the intervention was introduced. However, neither the proportion of early transmission nor the biological level of increased infectiousness was independently predictive of the long-term impact of treatment interventions. Both the long-term intervention impact and the amount of early transmission are dictated by the fundamental drivers of transmission—risk behaviors in the population and how they interact with the pathogens to determine $R_0$.

**Materials and Methods**

We developed a mathematical model of heterosexual HIV transmission in a two-sex population by using ordinary differential equations. The model incorporates heterosexual mixing between three sexual risk groups, a realistic representation of the natural history of HIV progression and infectiousness, and the effect of ART on survival and HIV infectiousness. Individuals may move among sexual risk groups, and sexual behavior may change over the course of the epidemic.

The model was calibrated to data about HIV prevalence and ART scale-up in South Africa by varying parameters controlling sexual behavior and infectiousness during early infection in a Bayesian statistical framework. This results in parameter combinations representing different underlying patterns of sexual behavior that are consistent with the observed HIV epidemic, allowing us to investigate how uncertainties about model parameters affect the impact of HIV treatment interventions. Full details and model equations are available in SI Appendix.

**Mathematical Model. Population structure and sexual mixing.** The model simulates a two-sex adult population aged 15 and older (Fig. 56). The population was divided into two age groups: 15–49, presumed to be a sexually active age group, and 50 and older, assumed not to form new sexual contacts but to possibly receive ART. Individuals enter the 15–49 population at a rate $\mu = 0.0226$ per year, move to the 50+ age group at a rate $\alpha = 1/35$ per year, and experience natural mortality from the age 50+ population at a rate $\mu = 1/11.45$

**Table 1. Estimates of early transmission, $R_0$, and intervention impact for fixed values for increased infectiousness during early infection**

<table>
<thead>
<tr>
<th>Relative infectiousness during early infection</th>
<th>Transmission during early infection when ART is introduced (2010), %</th>
<th>$R_0$ at epidemic start</th>
<th>During intervention period (after behavior change)</th>
<th>Reduction in HIV incidence rate after 31 y from an ART intervention (CD4 ≤350), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>No increase</td>
<td>1.9 (1.3–2.6)</td>
<td>8.2 (6.7–10.5)</td>
<td>4.1 (2.8–6.1)</td>
<td>21.4 (16.9–26.9)</td>
</tr>
<tr>
<td>9.2 times</td>
<td>11.4 (8.4–14.5)</td>
<td>5.8 (4.7–7.3)</td>
<td>3.6 (2.5–5.2)</td>
<td>21.8 (16.5–29.0)</td>
</tr>
<tr>
<td>26 times</td>
<td>18.8 (14.9–23.3)</td>
<td>4.6 (3.8–5.7)</td>
<td>3.5 (2.4–4.9)</td>
<td>21.2 (14.6–28.9)</td>
</tr>
</tbody>
</table>

Values in parentheses indicate 95% credible intervals.
per year, calibrated to match the composition of the South African population in 1990 and growth over time between 1990–2010 (63). The population was divided into HIV three sexual risk groups (termed “low,” “medium,” and “high”). As a crude means of simulating variability in sexual behavior over the life course, individuals may move from higher- to lower-risk groups at an annual rate ψ. Each sexual risk group has a different rate of forming new sexual contacts. A proportion of sexual contacts is reserved to be formed within the same sexual risk group according to an “assortativity” parameter ϵ (0.2, 0.8), and the remaining 1 − ϵ = rate of partnerships is formed randomly across the risk groups (49). The sexual contact rate in each risk group changed over time according to the logistic function

\[ \frac{1}{1 + \exp \left( - \left( t - \tau_c + d_c/2 \right) / d_t \right) } \]

where \( c_{rg}(t) \) is the contact rate for sex g and risk group r at time t, \( c_{rg}(0) \) is the contact rate at the start of the epidemic, \( \Delta_c \) is the overall proportion reduction in contact rate, \( \tau_c \) is the time behavior change begins, and \( d_t \) is the duration over which it occurs (Fig. 57). The rate of HIV transmission for contacts between susceptible and infected persons depends on the stage of HIV infection and ART status of the infected partner (see below) and partnership intensity parameter depending on the risk group of each partner. The proportion in each risk group, rate of movement among risk groups, relative contact rates, partnership intensity, and change in contact rate over time were varied in the model calibration.

HIV progression and infectiousness. HIV infection was divided into five stages (Fig. 58): early infection (mean 2.9 mo, CD4 > 350 cells/μL, 45.6 y), CD4 200–350 cells/μL (4.60 y), CD4 100–200 cells/μL (4.17 y), and CD4 ≤ 100 cells/μL (1.03 y to HIV death) (21), 50). Individuals progressed sequentially through these five stages before experiencing HIV mortality. The baseline HIV transmission rates per 100 person-years for the last four CD4 count stages are 4.4 (CD4 > 350), 7.2 (CD4 200–350), 27.1 (CD4 100–200), and 5.1 (CD4 ≤ 100) (2). For early infection, we defined a lognormal prior distribution for the relative infectiousness during early infection compared with CD4 > 350, with a mean of 26.0 times greater infectiousness and 95% prior mass between 12.6 and 47.8 (21).

ART model. Individuals may initiate ART from any stage of HIV infection. ART is divided into a multistage process (Fig. 59). Upon initiation, all individuals enter a “virtually suppression” stage during which the viral load is not yet suppressed (mean of 3 mo (51)). Infectiousness was assumed to be reduced by half during this stage compared with the CD4 stage from which treatment was initiated (Table S1). Following this stage, a proportion (0.189 for CD4 ≤ 100, 0.067 for CD4 100–200, 0.025 for CD4 200–350, 0.00 for CD4 > 350) moves to a “very sick” stage lasting for a mean of 6.2 mo before death, reflecting elevated mortality for patients starting treatment with low CD4 cell counts (52). The remainder of patients enter a long period of “effective ART,” during which infectiousness is reduced by 92% compared with untreated individuals with CD4 between 200 and 350 cells/μL (2). Individuals fail treatment (rate 0.03 per year for baseline CD4 ≤ 200, 0.026 for CD4 200–350, 0.022 for CD4 > 350) and enter a viremic stage in which they have the same infectiousness as persons with CD4 100–200 (mean 2.3 y), followed by a very sick stage before HIV death (mean 6.2 mo).

Persons on ART drop out at a rate of 0.12–0.168 per year, increasing with CD4 count at initiation (53) (Table S2), during the first 2 y and a rate of 0.088 thereafter (54). After dropping out from treatment, persons reenter a CD4 stage at or above that at which they initiated treatment, depending on their duration on treatment (Table S3), but progress at twice the rate of the treatment-naïve persons. They may reinitiate treatment once. See SI Appendix, section 1.5 for further details.

Statistical Methods. Data and likelihood. The model was calibrated to national HIV prevalence data among pregnant women attending antenatal care (ANC) from 1990 through 2008 (34), and HIV prevalence among men and women aged 15–49 y was estimated in nationally representative household surveys in 2002, 2005, and 2008 (55–57) (Fig. 1A). Prevalence from the household surveys was used as an unbiased estimate of true HIV prevalence for adult men and women. The difference between prevalence in women attending ANC and the general female population was described by an antenatal-bias parameter γ assumed to be constant over time on the logit scale, that is

\[ E(\logit(\gamma_{wm})) = \logit(\gamma_{ac}) + \gamma \]

where \( \gamma_{wm} \) is the HIV prevalence among all women aged 15–49 y at time t and \( \gamma_{ac} \) is the prevalence among women attending ANC. The likelihood was specified as a normal distribution around the logit-transformed prevalence estimates with the error variance estimated by the survey SE, accounting for the complex survey design of the ANC and household seroprevalence surveys (58). The influence of the existing ART scale-up in South Africa on prevalence and incidence was incorporated in the model calibration by simulating the percentage of adults (age 15+ years) on ART at midyear from 2005 through 2010 reported by the South African Department of Health (59) (Fig. S10). During this period, persons with CD4 ≤ 200 cells/μL were eligible to initiate ART. In the model calibration, persons with CD4 ≤ 100 cells/μL initiated ART at a rate eight times higher than those with CD4 100–200 cells/μL, and women initiated ART at rate 1.8 times higher than men to capture the median CD4 count and sex differential observed in patients initiating ART (60).

Estimated model parameters and prior distributions. Seventeen model parameters and the ANC bias parameter γ were estimated in the model calibration. Parameters related to the natural history of infection, the effects of ART, and demographics were fixed based on values from the literature as described above (Table S4). Parameters determining sexual behavior and sexual mixing, which are not easily relatable to directly observed measures, were estimated (Table S5). Estimated model parameters included the increased infectiousness during early infection relative to persons with CD4 > 350 cells per μL, proportion in each sexual risk group, mean sexual contact rate and relative contact rates for each risk group, group, degree of assortativity, rate of movement among risk groups, partnership transmission intensities, and start time of the epidemic (Table S5). The prior distribution for the relative infectiousness during early infection was lognormal (3.2, 0.34), resulting in a prior mean of 26.0 times increased infectiousness and 95% of the prior mass between 12.6 and 47.8, based on ref. 21. The assortativity parameter determining the proportion of partnerships formed exclusively within the same sexual risk group was restricted to between 0.2 and 0.8. The joint posterior distribution was estimated by using an incremental mixture importance sample (61). See SI Appendix, section 3.3, for further details; Fig. S1 for the marginal prior and posterior distributions of all parameters; and Table S6 for bivariate correlations between parameters in the posterior distribution.

\[ R_0 \text{ Calculation.} \] We calculated \( R_0(t) \) as the spectral radius of the next-generation matrix at time t (62). The next-generation matrix was calculated for a given set of parameter values by using the formalism described for compartmental systems in ref. 63, following which eigenvalues were solved numerically. Because the only time-varying parameter is the change in contact rate \( c_{rg}(t) \) (described above), the value of \( R_0 \) during the intervention period, after behavior change has occurred, may be expressed as \( R_0 \) at the start of the epidemic scaled by the proportion reduction in contact rate \( \Delta_c \) (see SI Appendix for details):

\[ R_0_{t_{start}} = R_0_{t_{start}} \left( 1 - \Delta_c \right) \]

The mathematical model was implemented in C++, and statistical calibration was implemented in C by using the GNU Scientific Library (64). Intervention analyses were conducted using R (65). Computer code to reproduce parameter estimation and model analyses is available for download from github.com/jeffreon/tasp-and-early-infection.

ACKNOWLEDGMENTS. We thank Prof. Geoffrey Garnett (Bill and Melinda Gates Foundation) for input to the development of the mathematical model. We thank Leigh Johnson (University of Cape Town) and Le Bao (Pennsylvania State University) for advice regarding data interpretation, statistical modeling, and computation. We thank Prof. Christophe Fraser (Imperial College London) for helpful discussions regarding the interpretation of the results. We thank Hannah Slater, Prof. Christophe Fraser, James Truscott, Steven Riley (Imperial College London), and Dédicre Hollingsworth (University College London) for assistance with the calculation and interpretation of \( R_0 \). We thank the Imperial College High Performance Computing Service (www.imperial.ac.uk/it/services/ hpc) for providing and maintaining computing resources. We thank the Bill and Melinda Gates Foundation for funding through a grant to the HIV Modeling Consortium. J.W.E. thanks the British Marshall Aid and Commemoration Commission for scholarship support.


Supporting Information

Eaton and Hallett 10.1073/pnas.1323007111

SI Appendix

The supplementary text provides a detailed description of the mathematical model. *SI1 Mathematical Model* gives a non-technical description of the model. *SI2 Model Equations* gives a mathematical description of the model. *SI3 Model Calibration* describes the calibration of the model to the South African epidemiologic data.

**SI1 Mathematical Model**

The model uses ordinary differential equations to simulate heterosexual HIV transmission in a two-sex population. The sexually active population is divided into three sexual risk groups that mix semiassortatively. Individuals may move among risk groups, and sexual behavior may change over the course of the epidemic. The model includes progression through five stages of HIV infection according to CD4 cell count, and allows for initiation of ART at any level of CD4 cell count.

1.1 Population Structure. The model simulates a two-sex adult population aged 15 and older (see diagram in Fig. S6). The population is divided into the age groups 15–49, which are presumed to be sexually active, and 50 and older, who are assumed to not form new sexual contacts. Individuals move from the younger to older age group at a rate \( \alpha = 1/35 \) per year and die from the age 50+ population at an annual rate \( \mu = 1/11.45 \), calibrated to match the relative sizes of the 15–49 year-old population and the age 50+ population in 1990 (1). Individuals enter the age 15–49 population at a rate \( \sigma = 0.0226 \) per year, calibrated such that the population growth over the period 1990–2010 approximately matches the population growth estimates over that period published by Statistics South Africa (1). All new sexual contacts, and hence HIV transmission, occur in the 15–49 age group, but the older age group is included in the model to assess the total ART need in the intervention scenarios. The sexually active population is divided into three sexual risk groups (termed “low,” “medium,” and “high”). As a crude means of simulating natural variability in individuals’ sexual risk behavior, individuals move from high risk to medium risk, high to low risk, and medium to low risk at a rate \( \psi \). This rate is varied in the model calibration (*SI3 Model Calibration*), and is the same for both sexes.

The proportion of the population in each risk group before the introduction of HIV into the population is allowed to be different for each sex and is estimated in the model calibration. The proportion of new entrants into the population who enter each risk group is calculated such that, in the absence of HIV, this proportion of the population in each risk group would remain constant. The proportion entering each risk group remains fixed over the duration of the simulation (meaning that the relative size of risk groups may change as a result of the differential burden of HIV in each risk group).

1.2 Sexual Mixing. As described in the previous section, the population is divided into three sexual risk groups (Fig. S6). The sexual contact rate in each of the risk groups is determined by the overall population average sexual contact rate, the relative rate of new sexual contacts among the three risk groups, and the size of the risk group. The underlying population average rate of new unprotected sexual contacts, \( \bar{c}(t) \), is allowed to vary over the course of the epidemic to model potential behavior change in response to the epidemic, such as increased condom use (2) or reductions in the number of new sexual partners (3). The functional form for the reduction is a logistic function parameterized by the initial contact rate, \( c_0 \); the proportion reduction in the contact rate that will occur, \( \Delta_c \); the start year of the behavior change, \( t_0 \); and the number duration (in years) over which the behavior change occurs, \( d_c \). Together, the overall average contact rate at time \( t \) is given by

\[
\bar{c}(t) = \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c}{1 + \exp\left(\frac{t - (t_0 + d_c/2)}{d_c/10}\right)} \quad \text{[S1]}
\]

The form of this logistic function ensures that the modeled behavior change is symmetric around the midpoint of the behavior change period \( t_0 + d_c/2 \) and that the change in behavior occurs in the interval \( t_0 \) to \( t_0 + d_c \). To see this, observe that \( \bar{c}(t_0) \approx c_0 \) and \( \bar{c}(t_0 + d_c) \approx c_0 \cdot (1 - \Delta_c) \):

\[
\bar{c}(t_0) = \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c}{1 + \exp\left(\frac{-d_c/2}{d_c/10}\right)} \quad \text{[S2]}
\]

\[
= \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c}{1 + \exp(-5)} \quad \text{[S3]}
\]

\[
= \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c \cdot 0.993}{1 + \exp(-5)} \quad \text{[S4]}
\]

\[
\approx \bar{c}_0 \quad \text{[S5]}
\]

and

\[
\bar{c}(t_0 + d_c) = \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c}{1 + \exp\left(\frac{t_0 + d_c - (t_0 + d_c/2)}{d_c/10}\right)} \quad \text{[S6]}
\]

\[
= \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c}{1 + \exp\left(\frac{d_c/2}{d_c/10}\right)} \quad \text{[S7]}
\]

\[
= \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c}{1 + \exp\left(\frac{d_c}{d_c/10}\right)} \quad \text{[S8]}
\]

\[
= \frac{c_0 \cdot (1 - \Delta_c) + \bar{c}_0 \cdot \Delta_c \cdot 0.993}{1 + \exp(5)} \quad \text{[S9]}
\]

\[
\approx \bar{c}_0 \cdot (1 - \Delta_c) \quad \text{[S10]}
\]

This is illustrated in the diagram in Fig. S7.

Each of these parameters (the initial contact rate, the percentage reduction in the contact rate, the timing of the start of the reduction, and the duration of the change) is estimated in the model calibration. The relative contact rates between high- and low-risk females and medium- and low-risk females also are estimated. The relative contact rates for males are calculated based on the relative contact rates for females and the sizes of each of...
the risk groups for males and females so that the total number of contacts offered by males and females in the same risk group is the same.

The number of sexual contacts formed between members of each risk group is determined by the sexual mixing parameter $e$, as proposed by Garnett and Anderson (4). A proportion $e$ (between 0 and 1) of sexual contacts is reserved exclusively to be formed with other members of the same risk group, whereas the remaining $(1 − e)$ proportion of partnerships is formed at random. The value of $e$ is estimated in the model calibration. If $e = 0$, sexual mixing is completely random, whereas if $e = 1$, mixing is fully assortative. As HIV mortality differentially affects males and females of each risk group, the total number of contacts offered by males may not be equal to the number of partnerships offered by females. In this case, the number of contacts desired by males and females is geometrically weighted by the parameter $\theta_r$, ranging between zero and one; $\theta_r = 0$ indicates that the females’ preferences determine the number of contacts, whereas $\theta_r = 1$ indicates that the males’ desired number of contacts dominates. For this exercise, the value is fixed at $\theta_r = 0.5$.

1.3 Natural History of HIV Infection. HIV infection is divided into stages according to the CD4 cell count progression associated with the duration of infection (Fig. S8). The stages are:

i) Primary infection

ii) CD4 count greater than 350 cells/µL

iii) CD4 count between 200 and 350 cells/µL

iv) CD4 count between 100 and 200 cells/µL and

v) CD4 count below 100 cells/µL

HIV-infected individuals progress from one stage of HIV infection to the next at a rate that is the reciprocal of the average duration in the stage. The average duration of primary infection is 2.9 mo, as estimated by Hollingsworth et al. (5). The rates of progression to subsequent CD4 cell stages and the overall duration from HIV infection to death are based on estimates of the progression to subsequent CD4 cell stages and the overall duration on treatment to simulate CD4 cell count strata in a collaborative analysis of ART cohorts from sub-Saharan Africa (11).

For most patients in whom ART is effective, the viral load is suppressed and transmission is reduced by 92% compared with the HIV transmission rate in the CD4 200–350 cells/µL stage (8). The period of effective ART is divided into two stages—first, a period of early effective ART lasting an average of 1.75 y, and then a long period of sustained viral suppression. The reduction in transmission is assumed to be the same in both stages, but this is implemented as separate ART stages so that the dropout rate from treatment can be varied according to the duration on ART. For example, we may assume that there is high dropout in the years following ART initiation, but patients who remain on treatment for 2 y likely have accommodated treatment and have high retention thereafter. In addition to the previously described higher probability for immediate treatment failure and death for those starting ART at low CD4 cell counts, the failure rate for long-term effective ART is assumed to be slightly lower for those who start treatment at high CD4 cell counts (Fig. S9), in line with observations that mortality is modestly higher even several years after treatment initiation for those who start at low CD4 cell counts (12) and to ensure there is no “survival benefit” in the model from delaying treatment initiation.

After patients fail long-term effective ART, they enter a stage of “treatment failure” in which they are viremic and are assumed to have the same infectiousness as individuals in the CD4 cell count category 200–350 cells/µL. The average duration of this stage is 2.3 y. Finally, individuals enter a stage of being very sick just before death, which lasts an average of 6.2 mo. During this period of being very sick, transmission is reduced and assumed to be at the same level as during the CD4 ≤100 cells/µL stage.

1.5 Dropping out from ART. Individuals may drop out from any of the first three stages of ART: virally suppressing, early effective ART, and effective ART. The rate of dropout from treatment is permitted to vary according to duration on ART and the CD4 cell count category from which treatment was initiated, in line with data suggesting that those starting treatment at higher CD4 cell counts may have poorer retention in treatment programs (13, 14). Rates of dropout from ART are representative of those observed in South African ART cohorts (12, 15–18). Assumed rates of dropping out from ART are presented in Table S2. Dropout is assumed to decline after the first 2 y on treatment, and individuals who reinitiate treatment after having dropped out once (see below) are assumed to have lower rates of dropout.

After dropping out from treatment, the untreated CD4 stage category that individuals enter depends on their pretreatment CD4 stage and the duration on treatment to simulate CD4 cell count reconstitution associated with ART. Individuals who drop out from the virally suppressing stage all return to the same CD4 stage from which they initiated treatment. For those who drop out
during the early effective ART stage, half move one CD4 stage. Those who drop out from the “effective ART” stage all increase two CD4 stages. This is summarized in Table S3. However, individuals who have dropped out of treatment progress through subsequent CD4 stages twice as fast as treatment-naive individuals (rates described in Fig. S9).

In the model, if individuals drop out of treatment, they are eligible to restart treatment again once. The rate at which these individuals reintegrate treatment depends on the CD4 cell count category, such that they are increasingly likely to reintegrate treatment at lower CD4 cell count categories, when they really are experiencing clinical symptoms. The rate of reintegrating treatment for those with CD4 cell counts between 200 and 350 cells/μL is 0.048 per year, the rate for those with CD4 cell counts between 100 and 200 cells/μL is 0.160, and the rate for those with CD4 counts below 100 cells/μL is 2.92 per year. Based on progression through these stages at twice the rate of treatment-naive individuals, the probability that an individual who dropped out will restart ART in each of these CD4 cell count categories before progressing to the next CD4 stage (or dying, in the case of those with CD4 counts ≤100 cells/μL) is 10%, 25%, and 60%, respectively. The overall probability that an individual who has dropped out of treatment once will reintegrate treatment before dying from HIV is 73%.

Upon restarting treatment, individuals progress through the same stages of ART as when first initiating ART. The baseline assumption for the rate of dropping out from treatment after reintegrating is 0.06 per year (Table S2, last column). Because individuals may restart treatment, but only once, the model separately tracks treated and untreated people according to the number of times they have initiated ART. To summarize this, Table S1 lists all the stages of antiretroviral treatment through which infected individuals may progress, and the subscript identifying each stage in the technical model description that follows.

1.6 HIV Transmission. The probability of transmission during contact between a susceptible and an infected individual depends on the annual transmission rate \( p_{m,r} \) in HIV stage \( m \) and treatment stage \( r \), the “intensity” of a contact \( \kappa_{u,r} \) between a male in risk group \( r_m \) and a female in risk group \( r_F \). The intensity parameter accounts for factors that affect the probability of transmission in different types of partnerships, such as the partnership duration, coital frequency, and condom use. The per-contact transmission probability based on these parameters is \( 1 - e^{-\kappa_{u,r} p_{m,r}} \), where \( m \) is the HIV stage of the infected partner.

1.7 Epidemic Seeding. We initialize the HIV epidemic with an adult HIV prevalence of 0.025% at time \( t_0 \) (estimated). This initial prevalence is distributed across the sexes, risk groups, and infected stages proportional to the eigenvector associated with the largest eigenvalue of the linearized Jacobian matrix describing transmission in a fully susceptible population (the matrix \( T + \Sigma \) defined in SI4 Rq. Calculation). Thus, the initial distribution of infections is consistent with that which would be expected during the early exponential growth period for a given set of parameters.

SI2 Model Equations

We divide the population into four categories according to their infection and treatment status:

- \( S^{g,r} \): HIV-uninfected and sexually active (susceptible) individuals of sex \( g \) in risk group \( r \).
- \( I^{g,r} \): HIV-infected and sexually active individuals of sex \( g \) in risk group \( r \) and HIV infection stage \( m \). The subscript \( u = 0 \) indicates untreated individuals who do not have access to treatment, \( u = 6 \) indicates individuals who have dropped out of treatment and are eligible to restart, and \( u = 12 \) indicates those who have dropped out of treatment a second time and are not eligible to restart.

\( T^{g,r}_{m,u} \): HIV-infected and sexually active individuals on ART (treated) who began treatment in HIV infection stage \( m \) and are in treatment stage \( u \) (Table S1).

\( R^{g,r}_{m,u} \): Individuals removed from the sexually active population of sex \( g \), in HIV infection stage \( m \), and treatment stage \( u \). (Uninfected individuals are indicated by \( m = 0 \) and untreated individuals by \( u \in \{0, 6, 12\} \).

In the above and throughout the following mathematical description, the superscript \( g \in \{M, F\} \) corresponds to sex; superscript \( r \in \{H, M, L\} \) corresponds to the sexual risk group for the sexually active population; subscript \( m \in \{0, \ldots, 5\} \) corresponds to HIV infection stage, with 0 representing uninfected and 1–5 corresponding to the stages of infection from primary infection to CD4 ≤100; and subscript \( u \in \{0, \ldots, 12\} \) corresponds to ART status, with 0 indicating untreated individuals without access to treatment, and the remaining stages indicating different stages of ART as indicated in Table S1.

The following differential equations define the dynamics of the groups:

\[
\frac{dS^{g,r}}{dt} = \alpha + \nu \frac{\pi^{g,r}}{2} S^{g,r+1} \left( I^{g,r+1} + T^{g,r+1} \right) + \sum_{r'} \psi_{r,r'} S^{g,r'}
\]

\[
- \left( \beta^{0,r} I^{g,r} \right) S^{g,r} + \sum_{r'} \psi_{r,r'} I^{g,r'} T^{g,r'}
\]

\[
\frac{dI^{g,r}_{1,0}}{dt} = \beta^{0,r} I^{g,r} S^{g,r} + \sum_{r'} \psi_{r,r'} I^{g,r'}_{1,0} - \left( \sigma_1 + \nu + \sum_{r'} \psi_{r,r'} \right) I^{g,r}_{1,0}
\]

\[
\frac{dI^{g,r}_{m,0}}{dt} = \beta^{0,r} I^{g,r}_{m-1,0} + \sum_{r'} \psi_{r,r'} I^{g,r}_{m,0} - \left( \sigma_1 + \nu + \sum_{r'} \psi_{r,r'} \right) I^{g,r}_{m,0}
\]

for \( m \geq 2 \)

\[
\frac{dT^{g,r}_{m,u}}{dt} = \lambda I^{g,r}_{m,0} + \sum_{r'} \psi_{r,r'} T^{g,r}_{m-1,u} - \left( \phi^{0,r}_{m,u} + \eta_{m,u} + \nu + \sum_{r'} \psi_{r,r'} \right) T^{g,r}_{m,u}
\]

\[
\frac{dT^{g,r}_{m,u}}{dt} = (1 - \frac{\pi^{g,r}}{2}) I^{g,r}_{m,1} + \sum_{r'} \psi_{r,r'} T^{g,r}_{m,u} - \left( \phi^{0,r}_{m,u} + \eta_{m,u} + \nu + \sum_{r'} \psi_{r,r'} \right) T^{g,r}_{m,u}
\]

\[
\frac{dT^{g,r}_{m,u}}{dt} = \phi^{0,r}_{m,u} T^{g,r}_{m,u-1} + \sum_{r'} \psi_{r,r'} T^{g,r}_{m,u} - \left( \phi^{0,r}_{m,u} + \eta_{m,u} + \nu + \sum_{r'} \psi_{r,r'} \right) T^{g,r}_{m,u}
\]

for \( u \in \{3, 4\} \)

\[
\frac{dT^{g,r}_{m,u}}{dt} = \psi^{0,r}_{m,1} T^{g,r}_{m,1} + \phi^{0,r}_{m,4} T^{g,r}_{m,4} + \sum_{r'} \psi_{r,r'} T^{g,r}_{m,u} - \left( \phi^{0,r}_{m,u} + \eta_{m,u} + \nu + \sum_{r'} \psi_{r,r'} \right) T^{g,r}_{m,u}
\]

for \( u \in \{5\} \)
\[
\frac{dR^g_{m,6}}{dt} = \sum_{m'} \sum_{u'=1}^{3} \beta_{m'u'} R^g_{m',u'} T^g_{m',u'} + \sigma_{m-1} R^g_{m-1,6} + \sum_r \psi_{r'} I^g_{m,6}
- \left( \sigma_m + \bar{\xi}_m + \nu + \sum_r \psi_{r'} \right) I^g_{m,6} \quad \text{for } m \geq 1
\]

\[
\frac{dT^g_{m,7}}{dt} = \lambda^g_m I^g_{m,6} + \sum_r \psi_{r'} T^g_{m,7} - \left( \phi^g_{m,7} + \eta_{m,7} + \nu + \sum_r \psi_{r'} \right) T^g_{m,7}
\]

\[
\frac{dT^g_{m,8}}{dt} = \left( 1 - \xi_m \right) \phi^g_{m,8} T^g_{m,8} + \sum_r \psi_{r'} T^g_{m,8}
- \left( \phi^g_{m,8} + \eta_{m,8} + \nu + \sum_r \psi_{r'} \right) T^g_{m,8}
\]

\[
\frac{dT^g_{m,9}}{dt} = \phi^g_{m,9} T^g_{m,9} + \sum_r \psi_{r'} T^g_{m,9}
- \left( \phi^g_{m,9} + \eta_{m,9} + \nu + \sum_r \psi_{r'} \right) T^g_{m,9}
\]

\[
\frac{dT^g_{m,11}}{dt} = \xi \phi^g_{m,11} T^g_{m,11} + \phi^g_{m,10} T^g_{m,10} + \sum_r \psi_{r'} T^g_{m,11}
- \left( \phi^g_{m,11} + \eta_{m,11} + \nu + \sum_r \psi_{r'} \right) T^g_{m,11}
\]

\[
\frac{dT^g_{m,12}}{dt} = \sum_{m'} \sum_{u'=6}^{9} \beta_{m'u'} T^g_{m',u'} + \sigma_{m-1} T^g_{m-1,12}
+ \sum_r \psi_{r'} T^g_{m,12}
- \left( \sigma_m + \bar{\xi}_m + \nu + \sum_r \psi_{r'} \right) T^g_{m,12}
\quad \text{for } m \geq 1
\]

\[
\frac{dR^g_{m,6}}{dt} = 3 \sum_{m'} \sum_{u'=1}^{3} \beta_{m'u'} R^g_{m',u'} R^g_{m'#u'} + \sigma_{m-1} R^g_{m-1,6}
+ \nu \sum_r R^g_{m,6} - \left( \sigma_m + \bar{\xi}_m + \mu \right) R^g_{m,6}
\quad \text{for } m \geq 1
\]

\[
\frac{dR^g_{m,7}}{dt} = \lambda^g_m R^g_{m,6} + \nu \sum_r T^g_{m,7} - \left( \phi^g_{m,7} + \eta_{m,7} + \mu \right) R^g_{m,7}
\]

\[
\frac{dR^g_{m,8}}{dt} = \left( 1 - \xi_m \right) \phi^g_{m,8} R^g_{m,7} + \nu \sum_r T^g_{m,8} - \left( \phi^g_{m,8} + \eta_{m,8} + \mu \right) R^g_{m,8}
\]

\[
\frac{dR^g_{m,11}}{dt} = \xi \phi^g_{m,11} R^g_{m,7} + \phi^g_{m,10} R^g_{m,10} + \nu \sum_r T^g_{m,11}
- \left( \phi^g_{m,11} + \eta_{m,11} + \mu \right) R^g_{m,11}
\quad \text{for } u \in \{9, 10\}
\]

\[
\frac{dR^g_{m,12}}{dt} = \sum_{m'} \sum_{u'=6}^{9} \beta_{m'u'} R^g_{m',u'} R^g_{m'#u'} + \sigma_{m-1} R^g_{m-1,12}
+ \nu \sum_r T^g_{m,12} - \left( \sigma_m + \bar{\xi}_m + \mu \right) R^g_{m,12}
\quad \text{for } m \geq 1
\]

In the above equations, the parameter \( \bar{\xi}^g_r \) is the proportion of new susceptible individuals of sex \( g \) that should enter risk group \( r \) to maintain a constant proportion \( \pi^g_r \) in risk group \( r \) in the absence of HIV infection. Solving the above equations with \( f^g_r(t) = 0 \) gives that

\[
\bar{\pi}^g_r = \frac{\sum_r \psi_{r'} \pi^g_{r'} - \sum_r \psi_{r'} \pi^g_{r'}}{\alpha + \nu}
\quad [S12]
\]

The function \( f^g_r(t) \) is the force of infection for the group \( S^g \) that depends on the contact rate \( c^g_r(t) \) at time \( t \) in that group, the probability that a contact is formed with an infectious partner, and the probability of transmission in that contact.

The contact rate for a risk group depends on the average underlying contact rate \( \bar{\tau}(t) \), which changes over time according to Eq. SI, the size of the risk groups at the beginning of the epidemic \( \pi^g_{r'} \), and the relative contact rates of the high- and medium-risk groups to the low-risk group \( \alpha^g_{r'} \) (where \( \alpha^g_{r'} := 1 \)). The weighted average of the relative contact rate by the size of the initial risk group yields the annual contact rate for each risk group at time \( t \):

\[
c^g_r(t) = \frac{\bar{\tau}(t) \cdot \alpha^g_r}{\sum_r \pi^g_r \alpha^g_r}
\]

The total number of contacts desired to be formed by members of risk group \( r \) of sex \( g \) thus is

\[
J^g_r(t) = c^g_r(t) \left( S^g + \sum_m I^g_{m,0} + \sum_m T^g_{m,0} \right)
\quad [S13]
\]

The number of these contacts \( J^g_r \) desired to be formed with each risk group \( r' \) of the opposite sex \( g' \) depends on the value of
the assortativity parameter $\varepsilon$. A proportion $\varepsilon$ of the partnerships are desired to be formed only with the members of the same risk group $r = r'$, whereas the remaining $1 - \varepsilon$ proportion of the partnerships are formed with each risk group of the opposite sex proportionally to the number of partnerships $J^{r'}$ offered by those risk groups. Formally, the proportion of contacts desired to be formed by sex $g$ and risk group $r$ that are formed with the risk group $r'$ of the opposite sex is defined as

$$Q^g_{r,r'} = \varepsilon \delta_{r,r'} + (1 - \varepsilon) \sum_r J^{r'}.$$  \[S14\]

where $\delta_{r,r'}$ is the Kronecker delta defined as $\delta_{r,r'} = 1$ if $r = r'$ and $0$ otherwise.

In the case that males in risk group $r_M$ and females in group $r_F$ do not agree on the number of partnerships to be formed between the risk groups, i.e., $Q^M_{r_M,r_F} + Q^F_{r_F,r_M} \neq 0$, the discrepancy is balanced according to the parameter $\theta_t$ as

$$\tilde{Q}^M_{r_M,r_F} = Q^M_{r_M,r_F} \left( \begin{array}{c} \phi^M_{r_M,r_F} \\ \phi^M_{r_F,r_M} \end{array} \right) \theta_t.$$  \[S15\]

The probability that transmission occurs in a contact between a susceptible and an infected individual depends on the stage of infection and treatment status of the infection according to the transmission rate parameter $\beta_{m,u}$ and on the value of the partnership intensity multiplier $\kappa_{u,t}$ for a partnership between the male in risk group $r_M$ and the female in risk group $r_F$. The force of infection then is calculated by summing over the probability that each contact is with an infectious individual and the probability that transmission occurs according to the infection stage and treatment status of the partner:

$$f^{r,r'}(t) = e^{d_{r,r'}(t)} \sum_r \left[ \frac{Q^r_{r,r'} \sum_m \sum_u T^{r'}_{m,u} p^{r'}_{m,u} + \sum_m T^{r'}_{m,u} p^{r'}_{m,u}}{S^{r,r'} + \sum_m \sum_u p^{r'}_{m,u} + \sum_m \sum_u T^{r'}_{m,u} p^{r'}_{m,u}} \right].$$  \[S16\]

where $p^{r,r'}_{m,u}$ is the probability of transmission per contact by an infected individual of sex $g$ in risk group $r$, HIV stage $m$, and treatment status $u$ to a susceptible individual of the opposite sex in risk group $r'$, defined as

$$p^{r,r'}_{m,u} = 1 - \exp\{-\beta_{m,u} \cdot \kappa_{u,t} \}.$$  \[S17\]

### SI3 Model Calibration

The model is calibrated to nationally representative HIV prevalence data and ART scale-up data from South Africa. The general strategy for model calibration is that parameters related to the natural history of HIV infection, the effect of ART on individual infection, and patterns of access and retention in the existing ART program are fixed and informed from the literature as described in SI1 Mathematical Model. Parameters relating to sexual behavior and mixing, the start time of the epidemic, the timing and magnitude of sexual behavior change, and the timing and rate of existing ART scale-up are estimated using a Bayesian approach. This yields a joint distribution of parameter combinations representing different sexual mixing patterns consistent with the observed epidemic.

#### 3.1 Data

The model is calibrated using HIV prevalence data from two sources. The first is HIV prevalence among 15–49-y-old pregnant women from the annual national antenatal prevalence surveys from 1990 to 2008 (19). The second is national HIV prevalence among 15–49-y-old males and females from the three nationally representative household surveys conducted by the Human Sciences Research Council in 2002, 2005, and 2008 (20–22). The discrepancy between the level of HIV prevalence in the antenatal surveillance and the household survey-based prevalence is reconciled by incorporating a bias parameter in the antenatal prevalence compared with prevalence among the general 15–49-y-old population from the household surveys. This is described in detail in section 3.3.

The model also is calibrated to the percentage of the adult population on ART. This is calculated by dividing the total number of adults reported to be on ART according to the South Africa Department of Health in June of each year from 2005 to 2010 (23) by the annual midyear population size estimate of those over 15 y old from Statistics South Africa (1). The resulting estimates for the proportion of the adult population on ART for 2005–2010 to which the model is calibrated are shown in Fig. S10.

#### 3.2 Estimated Model Parameters

A vector $\theta$ of 17 model parameters are estimated, and are used either directly or to derive a number of the model inputs in the equations described in SI2 Model Equations and Table S4. The mathematical model parameters that are estimated from fitting to HIV prevalence data from South Africa are given in Table S5, along with the prior probability distributions for the unknown parameters given the mathematical model and available HIV prevalence data, $W$. Let $\theta$ denote the set of parameters to be estimated from section 3.2 and $M$ denote the mathematical model and the associated fixed parameter values. For a given set of parameter values, the model produces a corresponding set of predicted HIV prevalence values $\xi(M(\theta))$. Then we specify a likelihood function $p(W|M(\theta))$ for the probability of the data given the model and parameter values. If we let $p(\theta)$ denote a prior distribution on the unknown parameters, using Bayes theorem the posterior distribution of $\theta$ given the model and data are given by

$$p(\theta|W,M) \propto p(\theta)p(W|M(\theta)).$$

#### 3.3 Statistical Methods

We use a Bayesian analysis to estimate probability distributions for the unknown parameters given the model and available HIV prevalence data, $W$. Let $\theta$ denote the set of parameters to be estimated from section 3.2 and $M$ denote the mathematical model and the associated fixed parameter values. For a given set of parameter values, the model produces a corresponding set of predicted HIV prevalence values $\xi(M(\theta))$. Then we specify a likelihood function $p(W|M(\theta))$ for the probability of the data given the model and parameter values. If we let $p(\theta)$ denote a prior distribution on the unknown parameters, using Bayes theorem the posterior distribution of $\theta$ given the model and data are given by

$$p(\theta|W,M) \propto p(\theta)p(W|M(\theta)).$$

##### 3.3.1 Likelihood function

We now derive the combined likelihood for the national seroprevalence survey data and antenatal clinic data. First, we define the likelihood for an individual datum.

For the national household survey estimates, let $W_{F,t}$ and $W_{M,t}$ be the HIV prevalence in the age group 15–49 reported by a survey in year $t$. As above, define $\xi_{g,t}$ to be the predicted population HIV prevalence for sex $g$ at time $t$ by the model $M(\theta)$. We assume that HIV prevalence estimates from national household surveys are an unbiased estimate of the true population prevalence, but we logit-transform the data to stabilize the error variance so that

$$\log\left( \frac{W_{F,t}}{1 - W_{F,t}} \right) = \log\left( \frac{\xi_{g,t}}{1 - \xi_{g,t}} \right) + \epsilon_{g,t},$$  \[S18\]

where $\epsilon_{g,t} \sim \text{Normal}(0, \sigma^2_{\epsilon,g,t})$ and the errors are conditionally independent given $\xi$. Thus, the likelihood function for an estimate from a national prevalence survey is
\[ p(W_{S,t}|M(\theta)) = \frac{1}{\sqrt{2\pi}\sigma_{S,t}} \exp \left\{ -\frac{1}{2\sigma_{S,t}^2} \left( \logit(W_{S,t}) - \logit(\xi_{S,t}) \right)^2 \right\}. \] [S19]

In calculating the likelihood, the value of \( \sigma_{S,t}^2 \) is replaced by an estimate \( \hat{\sigma}_{S,t}^2 \) based on the confidence intervals reported by the Human Sciences Research Council (HSRC) survey, which account for the complex sampling design. The confidence intervals in the HSRC reports are on the inverse-logit scale, so the error variances are estimated by

\[ \hat{\sigma}_{S,t}^2 = \left( \logit\left(\text{CI}_{\text{max}}^{S,t}\right) - \logit\left(\text{CI}_{\text{min}}^{S,t}\right) \right)^2, \] [S20]

where \( \Phi^{-1} \) is the inverse of the standard normal cumulative distribution function.

For the antenatal clinic data, similar to refs. 2 and 24, we assume that HIV prevalence \( W_{C,t} \) among antenatal clinic attendees at time \( t \) is linearly related to prevalence in the general female age 15–49 population on the logit scale and that this effect is fixed over time. To model this, we introduce an additional parameter \( \gamma \) such that

\[ \log\left( \frac{W_{C,t}}{1-W_{C,t}} \right) = \log(\frac{\xi_{C,t}}{1-\xi_{C,t}}) + \gamma + \epsilon_{C,t}, \] [S21]

where the error term \( \epsilon_{C,t} \sim \text{Normal}(0, \sigma_{C,t}^2) \) and recalling that \( \xi_{C,t} \) is the HIV prevalence among females aged 15–49 predicted by the model at time \( t \). Thus,

\[ p(W_{C,t}|M(\theta)) = \frac{1}{\sqrt{2\pi}\sigma_{C,t}} \exp \left\{ -\frac{1}{2\sigma_{C,t}^2} \left( \logit(W_{C,t}) - \logit(\xi_{C,t} + \gamma) \right)^2 \right\}. \] [S22]

Once again, when evaluating the likelihood, we replace \( \sigma_{C,t}^2 \) with an estimate \( \hat{\sigma}_{C,t}^2 \) calculated from the confidence intervals published by the South African Department of Health. In the case of the antenatal clinic data, the reported confidence intervals are symmetric, suggesting that they have been estimated on the untransformed scale rather than logit-transformed prevalence. Deriving an estimate of the sampling error variance on the logit scale involves two steps: first estimating the untransformed error variance \( \tau_{C,t}^2 \), and then using the delta method to approximate the variance \( \hat{\sigma}_{C,t}^2 \) of the logit-transformed distribution, which depends on the estimated antenatal clinic prevalence at time \( t \), \( W_{C,t} \). The equations for this are

\[ \hat{\tau}_{C,t}^2 = \left( \frac{\logit(W_{C,t}) - \logit(\xi_{C,t})}{2 \cdot \Phi^{-1}(0.975)} \right)^2 \] [S23]

\[ \hat{\sigma}_{C,t}^2 = \frac{\hat{\tau}_{C,t}^2}{W_{C,t}(1-W_{C,t})}. \] [S24]

To arrive at the likelihood for the full data \( W \), we assume that the data points are conditionally independent given the predicted prevalences \( \xi = M(\theta) \) and the antenatal clinic bias parameter \( \gamma \). Defining \( T_N \) to be the set of years for which national survey prevalence estimates are available and \( T_C \) the set of years for which antenatal clinic estimates are available, the full likelihood is

\[ p(W|M(\theta), \gamma) = \prod_{t \in T_N} \prod_{j \in [MF]} p(W_{S,t}|M(\theta)) \cdot \prod_{t \in T_C} p(W_{C,t}|M(\theta), \gamma) \]
\[ = \prod_{t \in T_N} \prod_{j \in [MF]} \frac{1}{\sqrt{2\pi}\sigma_{S,t}} \exp \left\{ -\frac{1}{2\sigma_{S,t}^2} \left( \logit(W_{S,t}) - \logit(\xi_{S,t}) \right)^2 \right\} \]
\[ \times \prod_{t \in T_C} \frac{1}{\sqrt{2\pi}\sigma_{C,t}} \exp \left\{ -\frac{1}{2\sigma_{C,t}^2} \left( \logit(W_{C,t}) - (\logit(\xi_{C,t} + \gamma)) \right)^2 \right\}. \] [S25]

### 3.3.2 Priors.

The prior distributions for the estimated model parameters \( \theta \) are given in Table S5. The antenatal bias parameter \( \gamma \) is assumed to have Uniform(0, 1) prior distribution. This amounts to assuming that the odds ratio of ANC prevalence to adult female prevalence is between 1 and 2.7.

### 3.3.3 Estimating the posterior distribution.

Multiplying the likelihood and prior distribution yields the joint posterior distribution of the parameters \( \theta \) and \( \gamma \) given the model and data up to a scaling constant:

\[ p(\theta, \gamma|W, M) \propto p(W|\theta, \gamma, M)p(\theta, \gamma). \] [S26]

We principally are interested in the values of the model parameters \( \theta \). The posterior distribution for \( \theta|W, M \) can be calculated by integrating out the parameter \( \gamma \). Observe that

\[ p(\theta|W, M) \propto \int_{\gamma} p(W|\theta, \gamma, M)p(\theta, \gamma)\,d\gamma \]
\[ = \int_{\gamma} \frac{p(\theta)p(\gamma)}{p(\theta)} p(W_{S,t}|M(\theta)) \cdot \prod_{t \in T_C} p(W_{C,t}|M(\theta), \gamma)\,d\gamma \]
\[ = \int_{\gamma} \prod_{t \in T_N} p(W_{S,t}|M(\theta))p(\theta) \cdot \prod_{t \in T_C} p(W_{C,t}|M(\theta), \gamma)\,d\gamma; \] [S27]

so if we can efficiently evaluate the integral \( \int_{\gamma} p(\gamma) \prod_{t \in T_C} p(W_{C,t}|M(\theta), \gamma)\,d\gamma \), then we can efficiently estimate the posterior distribution of \( \theta \). Before attacking this, let us define three useful quantities:

\[ S^2 = \left( \sum_{t \in T_C} \frac{1}{\sigma_{C,t}^2} \right)^{-1} \] [S28]

\[ \mathcal{T} = S^2 \cdot \sum_{t \in T_C} \frac{W_{C,t} - \xi_{C,t}}{\sigma_{C,t}^2} \] [S29]

\[ \mathcal{T}^2 = S^2 \cdot \sum_{t \in T_C} \frac{(W_{C,t} - \xi_{C,t})^2}{\sigma_{C,t}^2}. \] [S30]

The first may be thought of as the pooled variance of the ANC prevalence estimates, the second as the precision-weighted mean difference between ANC data prevalence and the model-predicted female prevalence, and the third as the precision-weighted mean-squared distance between the ANC data and the predicted female prevalence.

Now, again consider our integral. We will do a bit of rearranging to show that the integral can be evaluated as a normal cumulative...
distribution function. For brevity, denote \( \hat{W}_i = \logit(W_{ij}) \) and \( \zeta_i = \logit(\zeta_{ij}) \), and all sums and products are over the set \( T_C \).

\[
\int_{\hat{\Theta}} p(\gamma)p(W_C|M(0),\gamma)\,d\gamma = \prod_i \frac{1}{\sqrt{2\pi\sigma^2_i}} \int_0^{\infty} \exp\left\{-\frac{1}{2\sigma^2_i}(\hat{W}_i - (\zeta_i + \gamma))^2\right\} \,d\gamma
\]

\[
= \frac{1}{\prod_i 2\pi\sigma^2_i} \int_0^{\infty} \exp\left\{-\sum_i \frac{1}{2\sigma^2_i}(\gamma - (\hat{W}_i - \zeta_i))^2\right\} \,d\gamma
\]

\[
= K \cdot \exp\left\{-\frac{1}{2}\sum_i \left(\frac{\gamma^2}{\sigma^2_i} - \frac{2\lambda_i \hat{W}_i - \zeta_i}{\sigma^2_i} + \frac{(\hat{W}_i - \zeta_i)^2}{\sigma^2_i}\right)\right\}
\]

\[
= K \cdot \exp\left\{-\frac{1}{2}\left(\frac{\gamma^2}{\sigma^2} - 2\gamma \frac{W^0 - \zeta}{\sigma^0} + \frac{(W^0 - \zeta)^2}{\sigma^0}\right)\right\}
\]

\[
= K \cdot \exp\left\{-\frac{1}{2\sigma^2} \left(\frac{\gamma^2}{\sigma^2} + 2\frac{W^0 - \zeta}{\sigma^0} + \frac{(W^0 - \zeta)^2}{\sigma^0}\right)\right\}
\]

\[
= K \cdot e^{-\left(\frac{\gamma^2}{2\sigma^2}\right)} \left(\frac{\gamma^2}{2\sigma^2}\right)^{1/2} \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{\infty} e^{-\left(\frac{\gamma^2}{2\sigma^2}\right)} \,d\gamma
\]

\[
= \frac{2\pi\sigma^2}{\sqrt{2\pi\sigma^2}} \left(\frac{\gamma^2}{2\sigma^2}\right)^{1/2} \left(\frac{\gamma^2}{2\sigma^2}\right)^{-1/2} \left[1 = \Phi\left(\frac{\gamma}{\sqrt{\sigma^2}}\right)\right]
\]

\[
[831]
\]

Using this expression, we can efficiently evaluate the posterior density function \( p(\theta|W,M) \) up to a constant. We use the incremental mixture importance sampling algorithm to approximate and sample from the posterior distribution (25).

**S14 \( R_0 \) Calculation**

We calculate \( R_0(t) \) over the course of the epidemic as the dominant eigenvalue of the next-generation matrix (NGM) following the formalism for compartmental systems described by Diekmann et al. (26). Consider the Jacobian matrix representing the linearization of the infected subsystem (the equations \( R_{gg} \) defining the infected persons, excluding the susceptible \( S^g \), treated \( T^g_{m,n} \), and removed \( R^g_{m,n} \) stages). We decompose this into the sum of two matrices \( T + \Sigma \), where \( T \) describes transmissions giving rise to new infected persons and \( \Sigma \) describes all other transitions between infected states. We calculate \( R_0 \) as the dominant eigenvalue of the NGM with large domain, \( K_L = -T^{-1} \Sigma^{-1} \) (26).

Both \( T \) and \( \Sigma \) are \( 30 \times 30 \) matrices summarizing the rates of transmission and transition between \( g = 2 \) sexes, \( m = 5 \) stages of HIV infection, and \( r = 3 \) risk groups. The transmission matrix \( T_g \) gives the rate of new infections in state \( i \) created by an infected person in state \( j \). Because the contact rate \( \epsilon(r,t) \) varies over time (according to the parameter \( \Delta_G \); Eq. S1), \( T(t) \) is a function of time \( t \). Because all transmission occurs heterosexually, the transmission matrix \( T \) consists of 15 \( \times \) 15 submatrices.

\[
T(t) = \begin{bmatrix}
0 & T^{\rightarrow F} & T^{\rightarrow M} \\
T^{\leftarrow F} & 0 & 0 \\
T^{\leftarrow N} & 0 & 0
\end{bmatrix}
\]

[832]

All newly infected persons start in the first stage of infection (early transmission), and so the sex-specific transmission matrices consist of a row of \( 3 \times 3 \) submatrices \( T_g^{\rightarrow x} \) for the rate of transmission from infected persons in stage \( m \) with zeros below:

\[
T_g^{\rightarrow x}(t) = \begin{bmatrix}
T^g_{11} & T^g_{12} & T^g_{13} \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\]

\[
T_g^x_m(t) = \begin{bmatrix}
T^x_{11,m} & T^x_{12,m} & T^x_{13,m} \\
T^x_{21,m} & T^x_{22,m} & T^x_{23,m} \\
T^x_{31,m} & T^x_{32,m} & T^x_{33,m}
\end{bmatrix}
\]

[833]

The elements \( T_g^{\rightarrow x} \) are determined by the contact rate between risk groups \( r \) and \( r' \) of the opposite sex and the transmission probability per contact \( p_{g,rr'} \) between these risk groups for an infected person in stage \( m \) defined in Eq. S17:

\[
T_g^x_m(t) = \epsilon(r,t) \cdot \left( \delta_{r,r'} + (1 - \epsilon) \sum_{r''} \epsilon(r'',t) \pi_{rr''} \right) \cdot p_{g,rr'}
\]

[834]

The matrix \( \Sigma \) includes progression between disease stages (\( \sigma_m \)), movement from higher to lower sexual risk groups (\( \nu \)), and removal from the sexually active population (\( \nu = 1/35 \) per year). As these parameters are assumed fixed and the same for each sex, \( \Sigma \) does not depend on time and consists of identical 15 \( \times \) 15 submatrices on the block diagonal

\[
\Sigma = \begin{bmatrix}
\Sigma^G & 0 \\
0 & \Sigma^G
\end{bmatrix}
\]

[835]

The matrix \( \Sigma^G \) consists of \( 3 \times 3 \) submatrices \( \Sigma^G \) on the diagonal representing transitions between risk groups and removals from each disease stage, and scaled identity matrices on the subdiagonal for entrants into the next disease stage:

\[
\Sigma^G = \begin{bmatrix}
\Sigma^G & 0 & 0 & 0 & 0 \\
0 & \Sigma^G & 0 & 0 & 0 \\
0 & 0 & \Sigma^G & 0 & 0 \\
0 & 0 & 0 & \Sigma^G & 0 \\
0 & 0 & 0 & 0 & \Sigma^G
\end{bmatrix}
\]

[836]

where \( I_3 \) is the \( 3 \times 3 \) identity matrix and \( \Sigma^G_{m} = - (\nu + \sigma_m) I_3 + \Psi \) with \( \Psi = \begin{bmatrix} -2\nu & 0 & 0 \\
\Psi & -\Psi & 0 \\
\Psi & \Psi & 0
\end{bmatrix} \) representing transitions from higher to lower sexual risk groups.

We solve for \( R_0(t) \) numerically as the dominant eigenvalue of \( K_L = -T(t)^{-1} \Sigma^{-1} \). Note that because \( \epsilon(r,t) \) is scaled proportionally for all \( (g,r) \), it follows from Eq. S1 that

\[
T(t) = \begin{bmatrix}
1 - \Delta_G & 0 & 0 \\
0 & 1 - \Delta_G & 0 \\
0 & 0 & 1 - \Delta_G
\end{bmatrix}
\]

[837]

Thus,

\[
R_0(t) = \begin{bmatrix}
1 - \Delta_G & 0 & 0 \\
0 & 1 - \Delta_G & 0 \\
0 & 0 & 1 - \Delta_G
\end{bmatrix}
\]

[838]

and in particular, after simulated behavior change has completed at the start of the intervention period \( R_0(l_{12,2016}) = (1 - \Delta_G) R_0(0) \).
Note that following Diekmann et al., we can obtain the NGM $K$ from the NGM with large domain $K_L$ by $K = E'K_LE$, where

$$E' = \begin{bmatrix} I_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \end{bmatrix}.$$


Fig. S1. Posterior density estimate of model parameters. Solid black densities indicate the posterior density, and dashed red densities indicate the prior density of each parameter. Solid vertical green lines indicate the posterior median, and dashed vertical blue lines indicate the bounds of posterior 95% credible intervals. Parameters are defined in Table S5.
Fig. S2. The relationship between (A) early transmission in 2010, (B) $R_0$ at the start of the epidemic, and (C) $R_0$ during the intervention period and model parameters: log of relative infectiousness during early infection (Left; $\beta_E/\beta_A$ in Table S5), the rate of movement from higher- to lower-risk groups (Center; $\psi$ in Table S5), and the percentage reduction in sexual contact rate over time (Bottom right; $\Delta_c$ in Table S5). The leftmost plot indicates the relationship between the outcome and relative early infectiousness ($\beta_E/\beta_A$). Center and Right indicate the relationship between the residual variation in the outcome (after removing the effect of $\beta_E/\beta_A$) and $\psi$ (Center) or $\Delta_c$ (Right). Residual $R^2$ values indicate the fraction of residual variance explained by $\psi$ (Center) and $\Delta_c$ (Right), and combined $R^2$ values indicate the fraction of all variance explained by early infectiousness and the plotted parameter together. All three parameters together explain 95% of the variation in $R_0$ during the intervention period (C).
Fig. S3. Sensitivity of results to assumptions about ART eligibility, coverage, and scale-up. In each panel, the black line indicates the baseline intervention strategy with eligibility for CD4 ≤350 cells/μL and 80% accessing treatment at a fixed rate after becoming eligible (illustrated in Fig. 3). (Left) Posterior mean reduction in HIV incidence rate over time; (Right) the correlation between the reduction in incidence and the percentage of transmission during primary infection at the start of the intervention in 2010 (analogous to Fig. 3B). (A) Changing the threshold for ART eligibility to all HIV-positive adults or only those with CD4 ≤200 cells/μL. All assume that eligible persons initiate treatment at a constant rate such that 80% will access treatment before dying from HIV. (B) Varying the percentage of persons who will access treatment. All assume eligibility for those with CD4 ≤350 cells/μL. (C) Changing the assumption about timing of ART initiation from assuming that all eligible persons will initiate treatment at a constant rate to assuming that those who will access treatment initiate it, on average, 1 y after becoming eligible. Both assume 80% of HIV-positive persons will access treatment and eligibility for those CD4 ≤350 cells/μL.
Fig. S4. Model calibration assuming persons in early infection are (i) no more infectious, (ii) 9.2 times more infectious, and (iii) 26 times more infectious than asymptomatic HIV-infected persons. (A) Posterior model fit to South Africa HIV prevalence data. (B) Posterior distribution of HIV incidence rate. (C) Posterior mean of the percentage of transmissions resulting from each stage of infection over time.
Fig. S5. Posterior density estimate for model parameters conditional on a fixed value for the increased infectiousness during early infection ($\beta_T/\beta_A$): no more infectious (green), 9.2 times more infectious (red), and 26 times more infectious (blue). Gray lines indicate the posterior densities for the full posterior distribution (Fig. S1).

Fig. S6. Diagram of model population structure and sexual mixing. Solid arrows indicate flows of individuals among risk groups. Dashed light gray arrows indicate sexual contacts among risk groups. Contacts are assortative such that individuals are more likely to form contacts with those in the same risk group.
Fig. S7. Logistic curve describing behavior change in Eq. S1.

Fig. S8. Average duration of and relative HIV transmission rate during each stage of HIV infection. (Note that vertical axis is not drawn to scale.)

Fig. S9. Stages of ART. Arrows indicate possible movements of individuals between stages. Black arrows indicate natural progression of individuals. Red arrows indicate dropout from ART. Individuals may drop out from ART only once and may return to a different CD4 stage depending on their duration on ART (SI Appendix, Tables S4 and S5).
Fig. S10. The percentage of adults (aged 15+) on ART over time used in the model calibration, based on reported numbers on ART from the South Africa Department of Health and population size estimates from Statistics South Africa.

Table S1. Stages of ART

<table>
<thead>
<tr>
<th>Subscript</th>
<th>Stage</th>
<th>Duration</th>
<th>Infectiousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Untreated, ART naïve</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Virally suppressing</td>
<td>3 mo</td>
<td>50% lower than prev. stage</td>
</tr>
<tr>
<td>2</td>
<td>Early effective ART, virally suppressed</td>
<td>1.75 y</td>
<td>92% lower than CD4 ≤350</td>
</tr>
<tr>
<td>3</td>
<td>Effective ART, virally suppressed</td>
<td>(Fig. S9)</td>
<td>92% lower than CD4 ≤350</td>
</tr>
<tr>
<td>4</td>
<td>Treatment failing, viremic</td>
<td>2.3 y</td>
<td>Same as CD4 200–350</td>
</tr>
<tr>
<td>5</td>
<td>Very sick</td>
<td>6.2 mo</td>
<td>Same as CD4 ≤100</td>
</tr>
<tr>
<td>6</td>
<td>Untreated, dropped out after first initiation, eligible to restart ART</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Reinitiated ART, virally suppressing</td>
<td>3 mo</td>
<td>50% lower than prev. stage</td>
</tr>
<tr>
<td>8</td>
<td>Reinitiated ART, early effective ART</td>
<td>1.75 y</td>
<td>92% lower than CD4 ≤350</td>
</tr>
<tr>
<td>9</td>
<td>Reinitiated ART, effective ART</td>
<td>(Fig. S9)</td>
<td>92% lower than CD4 ≤350</td>
</tr>
<tr>
<td>10</td>
<td>Reinitiated ART, treatment failing, viremic</td>
<td>2.3 y</td>
<td>Same as CD4 100–200</td>
</tr>
<tr>
<td>11</td>
<td>Reinitiated ART, very sick</td>
<td>6.2 mo</td>
<td>Same as CD4 ≤100</td>
</tr>
<tr>
<td>12</td>
<td>Untreated, dropped out after second initiation, not eligible to restart</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S2. Rate per year of dropping out from ART

<table>
<thead>
<tr>
<th>Baseline CD4 cell count</th>
<th>Duration on treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Virally suppressing</td>
</tr>
<tr>
<td>&gt;350</td>
<td>0.168</td>
</tr>
<tr>
<td>200–350</td>
<td>0.168</td>
</tr>
<tr>
<td>100–200</td>
<td>0.156</td>
</tr>
<tr>
<td>≤100</td>
<td>0.120</td>
</tr>
</tbody>
</table>
Table S3. CD4 stage after dropping out of treatment

<table>
<thead>
<tr>
<th>CD4 stage at ART initiation</th>
<th>Treatment stage at dropout</th>
<th>Dropouts returning to CD4 category, %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CD4 &gt;350</td>
</tr>
<tr>
<td>&gt;350</td>
<td>Virally suppressing</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Early effective ART</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Effective ART</td>
<td>100</td>
</tr>
<tr>
<td>200–350</td>
<td>Virally suppressing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Early effective ART</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Effective ART</td>
<td>100</td>
</tr>
<tr>
<td>100–200</td>
<td>Virally suppressing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Early effective ART</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Effective ART</td>
<td>100</td>
</tr>
<tr>
<td>≤100</td>
<td>Virally suppressing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Early effective ART</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Effective ART</td>
<td>100</td>
</tr>
</tbody>
</table>

Table S4. Model parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>Population growth rate (in absence of HIV)</td>
<td>0.023 per year</td>
</tr>
<tr>
<td>ν</td>
<td>Rate of progression from 15–49 to 50+ age groups</td>
<td>1/35 per year</td>
</tr>
<tr>
<td>μ</td>
<td>Mortality rate out of the 50+ age group</td>
<td>1/11.45 per year</td>
</tr>
<tr>
<td>π&lt;sub&gt;g,r&lt;/sub&gt;</td>
<td>Proportion of the age 15–49 population of sex g in risk group r in absence of HIV</td>
<td>Estimated</td>
</tr>
<tr>
<td>π&lt;sub&gt;g,r'&lt;/sub&gt;</td>
<td>Proportion of new entrants of sex g entering risk group r</td>
<td>Derived from π and ψ</td>
</tr>
<tr>
<td>ψ&lt;sub&gt;r&lt;/sub&gt;</td>
<td>Annual rate of moving from risk group r to r' for sex g</td>
<td>Estimated</td>
</tr>
<tr>
<td>ξ(t)</td>
<td>Population mean contact rate per year at time t</td>
<td>Estimated</td>
</tr>
<tr>
<td>ω&lt;sub&gt;g,r'&lt;/sub&gt;</td>
<td>Relative contact rate between risk group r and low-risk group for sex g</td>
<td>Estimated</td>
</tr>
<tr>
<td>ε</td>
<td>Degree of assortative mixing</td>
<td>Estimated</td>
</tr>
<tr>
<td>θ&lt;sub&gt;G&lt;/sub&gt;</td>
<td>Balance between male and female partner preference</td>
<td>0.5</td>
</tr>
<tr>
<td>κ&lt;sub&gt;rM,rF&lt;/sub&gt;</td>
<td>Intensity of partnership between male in risk group r&lt;sub&gt;M&lt;/sub&gt; and female in r&lt;sub&gt;F&lt;/sub&gt;</td>
<td>Estimated</td>
</tr>
<tr>
<td>β&lt;sub&gt;m,u&lt;/sub&gt;</td>
<td>Annual HIV transmission rate in stage m and ART status u</td>
<td>See Figs. S8 and S9</td>
</tr>
<tr>
<td>α&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Rate of progression from HIV stage m to stage m + 1</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.94</td>
</tr>
<tr>
<td>δ&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Rate of progression from HIV stage m to stage m + 1 after treatment dropout</td>
<td>2π</td>
</tr>
<tr>
<td>λ&lt;sub&gt;g&lt;/sub&gt;&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Rate of ART initiation for sex g in stage m</td>
<td>See text</td>
</tr>
<tr>
<td>λ&lt;sub&gt;g&lt;/sub&gt;&lt;sub&gt;m'&lt;/sub&gt;</td>
<td>Rate of reinitiating ART after dropout for sex g in stage m</td>
<td>SI Appendix, section 1.5</td>
</tr>
<tr>
<td>φ&lt;sub&gt;m,u&lt;/sub&gt;</td>
<td>Rate of progression from ART stage u to u + 1 when initiating ART in HIV stage m for sex g</td>
<td>Fig. S9</td>
</tr>
<tr>
<td>ξ&lt;sub&gt;m&lt;/sub&gt;</td>
<td>Probability of immediate treatment failure if initiating ART in stage m</td>
<td>[0.025, 0.067, 0.189]</td>
</tr>
<tr>
<td>η&lt;sub&gt;m,u&lt;/sub&gt;</td>
<td>Rate of dropping out of ART if initiated in stage m and currently in stage u</td>
<td>Table S2</td>
</tr>
<tr>
<td>η&lt;sub&gt;m,u'&lt;/sub&gt;</td>
<td>Rate of dropping out of ART after reinitiating if reinitiated in stage m and currently in stage u</td>
<td>Table S2</td>
</tr>
<tr>
<td>ρ&lt;sub&gt;m,u&lt;/sub&gt;</td>
<td>Probability of entering CD4 stage m after dropping out of stage (m',u')</td>
<td>Table S3</td>
</tr>
<tr>
<td>t&lt;sub&gt;0&lt;/sub&gt;</td>
<td>Date at which HIV epidemic is seeded into population</td>
<td>Estimated</td>
</tr>
<tr>
<td>δ&lt;sub&gt;m, r&lt;/sub&gt;</td>
<td>Seed HIV prevalence for sex g, risk group r, and disease stage m</td>
<td>SI Appendix, section 1.7</td>
</tr>
<tr>
<td>Parameter</td>
<td>Description</td>
<td>Prior</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>$t_0$</td>
<td>Start date of the epidemic</td>
<td>Unif(1983, 1988)</td>
</tr>
<tr>
<td>$1 - \omega^{M,L}$</td>
<td>Proportion of males not in the low-risk group</td>
<td>Unif(0.05, 0.7)</td>
</tr>
<tr>
<td>$1 - \omega^{F,L}$</td>
<td>Proportion of females not in the low-risk group</td>
<td>Unif(0.05, 0.7)</td>
</tr>
<tr>
<td>$\pi_M$</td>
<td>Proportion of males in high-risk group of those not in low-risk group</td>
<td>Unif(0.2, 0.8)</td>
</tr>
<tr>
<td>$\pi_F$</td>
<td>Proportion of females in high-risk group of those not in low-risk group</td>
<td>Unif(0.2, 0.8)</td>
</tr>
<tr>
<td>$\Psi$</td>
<td>Annual rate of movement from higher- to lower-risk groups</td>
<td>Unif(0.0, 0.15)</td>
</tr>
<tr>
<td>$\zeta_0$</td>
<td>Mean annual contact rate at start of the epidemic</td>
<td>Unif(0.5, 4.0)</td>
</tr>
<tr>
<td>$\Delta_c$</td>
<td>Proportion reduction in average contact rate</td>
<td>Unif(0.0, 0.7)</td>
</tr>
<tr>
<td>$t_r + \Delta_c$</td>
<td>Year behavior change starts</td>
<td>Unif(1990, 2002)</td>
</tr>
<tr>
<td>$\omega^{F,M}$</td>
<td>Year behavior change ends</td>
<td>Unif(2002, 2010)</td>
</tr>
<tr>
<td>$\omega^{F,H} - \omega^{F,M}$</td>
<td>Relative contact rate between medium-risk and low-risk females</td>
<td>Unif(1, 70)</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Assortativity of sexual mixing</td>
<td>Unif(0, 0.5)</td>
</tr>
<tr>
<td>$\kappa_M$</td>
<td>Partnership intensity for partnership involving a high-risk partner</td>
<td>Unif(0, 0.2)</td>
</tr>
<tr>
<td>$\kappa_M$</td>
<td>Partnership intensity for partnership between medium and low risk</td>
<td>*</td>
</tr>
<tr>
<td>$\kappa_L$</td>
<td>Partnership intensity for partnership between low-risk partners</td>
<td>*</td>
</tr>
<tr>
<td>$\beta_E/\beta_A$</td>
<td>Ratio of infectiousness during early and asymptomatic (CD4 &gt;350) infection</td>
<td>LogNormal(3.2, 0.34)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Bias in ANC prevalence and 15–49-y-old female prevalence on the logit scale</td>
<td>Unif(0.0, 1.0)</td>
</tr>
</tbody>
</table>

*These parameters have a joint uniform prior distribution such that $0 < \kappa_L < \kappa_M < \kappa_H < 1.$

### Table S6. Bivariate correlations between posterior parameter estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$t_0$</th>
<th>Risk group size</th>
<th>$\Psi$</th>
<th>Contact rate</th>
<th>$\omega$</th>
<th>$\epsilon$</th>
<th>Partnership intensity</th>
<th>Early infectious</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_0$</td>
<td>-0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.3</td>
<td>-0.1</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.0</td>
</tr>
<tr>
<td>$1 - \pi^{M,L}$</td>
<td>-0.0</td>
<td>0.7</td>
<td>0.5</td>
<td>-0.3</td>
<td>0.3</td>
<td>0.1</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>$1 - \pi^{F,L}$</td>
<td>0.1</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>$\pi^{M,H}/(1 - \pi^{M,L})$</td>
<td>0.0</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.0</td>
</tr>
<tr>
<td>$\pi^{F,H}/(1 - \pi^{F,L})$</td>
<td>0.0</td>
<td>-0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>-0.0</td>
</tr>
<tr>
<td>$\Psi$</td>
<td>0.3</td>
<td>0.3</td>
<td>0.5</td>
<td>-0.0</td>
<td>0.0</td>
<td>-0.2</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>$\zeta_0$</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>-0.0</td>
<td>0.2</td>
<td>-0.2</td>
<td>0.0</td>
<td>-0.0</td>
</tr>
<tr>
<td>$\Delta_c$</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>-0.0</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>$\omega$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.0</td>
<td>0.0</td>
<td>0.2</td>
<td>-0.2</td>
<td>-0.0</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>0.0</td>
<td>-0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>-0.0</td>
</tr>
<tr>
<td>$\kappa_M$</td>
<td>0.1</td>
<td>-0.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>-0.6</td>
<td>0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>$\kappa_L$</td>
<td>-0.1</td>
<td>0.0</td>
<td>0.2</td>
<td>-0.6</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>-0.0</td>
</tr>
<tr>
<td>$\beta_E/\beta_A$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>-0.1</td>
<td>0.0</td>
<td>-0.5</td>
<td>0.1</td>
<td>-0.3</td>
</tr>
</tbody>
</table>

Parameters defined in Table S5.