Predicting the outcomes of treatment to eradicate the latent reservoir for HIV-1

Alison L. Hill, Daniel I. S. Rosenbloom, Feng Fu, Martin A. Nowak, and Robert F. Siliciano

The latent reservoir (LR) for HIV-1 is a population of long-lived resting memory CD4+ T cells with integrated HIV-1 DNA. After establishment during acute infection, it increases to \(10^5\) to \(10^7\) cells and then remains stable. As only T cells (Fig. 1A) HIV cure of latency is supported by case studies of stem cell transplantation (22, 23) and, more recently, early treatment initiation (24, 25), which have allowed patients to interrupt treatment for months or years, and, more recently, early treatment initiation (24, 25), which have allowed patients to interrupt treatment for months or years. Greater than \(10,000\)-fold reductions in reservoir size are required to permit a majority of patients to interrupt treatment before viral rebound occurs (Fig. 1B). A latent infected cell can produce virions, resulting in the dramatic reductions in reservoir size accompanying these strategies stands in stark contrast to the actions of current LRAs, which induce only a fraction of latent virus in vitro (26, 27) and have not produced a measurable decrease in LR size in vivo (12, 13, 28). It is unclear how patient outcomes depend on reservoir reduction between these extremes, nor even whether a reduction that falls short of those achieved with stem cell transplantation will bring any clinical benefit. LRA research needs to address the question: How low must we go?

In the absence of clinical data, mechanistic mathematical models can serve as a framework to predict results of novel interventions and plan clinical trials. When results do become available, the models can be tested and refined. Mathematical models have a long tradition of informing HIV-1 research and have been particularly useful in understanding HIV-1 treatment. Previous models have explained the multiphasic decay of viremia during ART (29), the initial seeding of the LR during acute infection (30), the limited inflow to the LR during treatment (31), the dynamics of viral blips (32), and the contributions of the LR to drug resistance (33). No model has yet been offered to describe the effect of LRAs. Here we present a novel modeling framework to predict the degree of reservoir reduction needed to prevent viral rebound following ART interruption. The model can be used to estimate the probability that cure is achieved, or, barring that outcome, to estimate the length of time following treatment interruption before viral rebound occurs (Fig. 1A).

**Results**

**Determination of Key Viral Dynamic Parameters Governing Patient Outcomes.** We use a stochastic model of HIV-1 reservoir dynamics and rebound that, in its simplest form, tracks two cell types: productively infected activated CD4+ T cells and latently infected resting CD4+ T cells (Fig. 1B). A latently infected cell can either activate or die, each with a particular rate constant. An actively infected cell can produce virions, resulting in the

**Significance**

HIV infection cannot be cured by current antiretroviral drugs, due to the presence of long-lived latently infected cells. New antilatency drugs are being tested in clinical trials, but major unknowns remain. It is unclear how much latent virus must be eliminated for a cure, which remains difficult to answer empirically due to few case studies and limited sensitivity of viral reservoir assays. In this paper, we introduce a mathematical model of HIV dynamics to calculate the likelihood and timing of viral rebound following antilatency treatment. We derive predictions for the required efficacy of antilatency drugs, and demonstrate that rebound times may be highly variable and occur after years of remission. These results will aid in designing and interpreting HIV cure studies.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1. A.L.H. and D.I.S.R. contributed equally to this work.

2. To whom correspondence should be addressed. Email: alhill@fas.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1406663111/-/DCSupplemental.

**PNAS Early Edition | 1 of 6**
and SI Materials and Methods
≈ \[1\]
P \delta
A
qAP
200 copies per mL. We computed the relationship

\[
P_{\text{Clr}}(q) \approx e^{-4/\delta qAP}.
\]

The expression \(qAP/\delta\) approximates the expected number of
fated-to-establish cells that will ever exit from the LR, explaining
the Poisson form of this expression. In SI Materials and Methods,
we provide the full derivation, as well as a formula—[SB]—for
the probability that rebound occurs a given number of days
following treatment interruption (a function of \(\delta, A, P_{\text{Est}}, r,\)
and efficacy \(q\)). Of note, the initial size of the reservoir itself is not
included among these parameters: although it factors into both \(A\)
(the product of the pre-LRA reservoir size and the per-cell ac-
tivation rate), and \(q\) (the ratio of post-LRA to pre-LRA reservoir
size), it does not independently influence outcomes. Both of
these formulas provide an excellent match to the explicit simu-
lation of the model (Fig. 2). The key assumption required for the
analysis is that \(r\) greatly exceeds \(\delta\); because viral doubling times
during rebound are measured on the order of a few days,
whereas LR decay is measured on the order of many months
or years, this assumption is expected to hold. Likelihood-based
inference can therefore proceed by efficient computation of re-
bound probabilities (using [SB]), rather than by time-consuming
stochastic simulation.

Outcomes depend only on the four parameters above even in
more complex models of viral dynamics, including other features
of T-cell biology and the HIV lifecycle (SI Materials and Meth-
ods). Alternate models studied include explicit tracking of free
virus with varying burst sizes, an eclipse phase during which an
infected cell produces no virus, proliferation of cells upon reac-
tivation, maintenance of the LR by homeostatic proliferation,
and either a constant or Poisson-distributed number of infected
cells produced by each cell (Figs. S2–S7). If proliferation of
latently infected cells is subject to high variability, e.g., by bursts
of proliferation, then rebound time and cure probability increase
slightly beyond the predictions of the basic model (Figs. S6 and S7).
No other modification to the model altered outcomes. Outcomes
of LRA therapy therefore are likely to be insensitive to details of
the viral lifecycle; accordingly, few parameters must be estimated
to predict outcomes.

Predicted Prospects for Eradicating Infection or Delaying Time to
Rebound. Using best estimates of parameters derived from pre-
viously reported data (Table 1), we can explore the likely out-
comes of interventions that reduce the LR. The best outcome of
LRA therapy, short of complete and immediate eradication, is
that so few latently infected cells survive that none reactivate and
start a resurgent infection during the patient’s lifespan. In this
case, LRA has essentially cleared the infection and cure is
achieved. We simulated the model to predict the relationship
between LRA efficacy and clearance (Fig. 2A). We find that
the LR must be reduced 10,000-fold before half of patients are
predicted to clear the infection.

If LRA therapy fails to clear the infection, the next-best out-
come is extension of the time until rebound, defined as plasma
HIV-1 RNA > 200 copies per mL. We computed the relationship
between LRA efficacy and median time until rebound among
patients who do not clear the infection (Fig. 2B). Roughly a 2,000-
fold reduction in LR size is needed for median rebound times of
1 y. Only modest (about 2-fold) increases in median rebound
time are predicted for up to 100-fold reductions in LR size. In this
range, the rebound time is independent of latent cell lifespan
(decay rate \(\delta\)) and is driven mainly by the reactivation rate (\(A\))
and the infection growth rate (\(r\)). The curve inflects upward (on a log
scale) at \(~100\)-fold reduction and eventually reaches a ceiling as

active infection of some number of other cells, or it can die from
other causes without producing virions that infect other cells. In
the latter case, cytotoxic T lymphocyte (CTL) killing, errors in
viral reverse transcription, or other problems upstream of virion
production may prevent further infection. The model only tracks
the initial stages of viral rebound, when target cells are not yet
limited. A full description of the model is provided in Materials and
Methods and SI Materials and Methods.

The initial conditions for the dynamic model depend on the
number of latently infected cells left after LRA therapy. LRA
efficacy is defined by the fraction \(q\) of the LR that remains fol-
lowing treatment. The model tracks each latent and active cell to
determine whether viral rebound occurs, and if so, how long it
takes. Importantly, no single activated cell is guaranteed to
reactivate the infection, as it may die before infecting other cells.
Even if it does infect others, those cells likewise may die
before completing further infection. This possibility is a general
property of stochastic models, and the specific value for the
establishment probability depends on the rates at which infection
and death events occur. Our goal is to calculate the probability
that at least one of the infected cells remaining after therapy
escapes extinction and causes viral rebound, and if so, how long
it takes. If all cells die, then rebound never occurs and a cure is
achieved. As the model only describes events after completion
of LRA therapy, our results are independent of the therapy pro-
tocol or mechanism of action.

Using both stochastic simulations and theoretical analysis of
this model, we find that the probability and timing of rebound
relies on four key parameters: the decay rate of the LR in
the absence of viral replication (\(\delta\)), the rate at which the LR pro-
duces actively infected cells (\(A\)), the probability that any one
activated cell will produce a rebounding infection before its
lineage dies (\(P_{\text{Est}}\)), and the net growth rate of the infection once
restarted (\(r\)). Estimates of these four parameters are provided in
Table 1 and Fig. S1. After therapy, the rate at which the LR
produces actively infected cells is reduced to \(qA\). The probability
that an individual successfully clears the infection is

\[
P_{\text{Clr}}(q) = e^{-4/\delta qAP}.
\]
clearance of the infection becomes the dominant outcome. The upward inflection results from a change in the forces governing viral dynamics. If the reservoir is large (little reduction), then cells activate frequently, and the dominant component of rebound time is the time that it takes for virus from the many available activated cells to grow exponentially to rebound levels; the system is in a growth-limited regime. If the reservoir is small (large reduction), the dominant component is instead the expected waiting time until activation of the first cell fated to establish a rebounding lineage; the system is in an activation-limited regime. Because waiting time is roughly exponentially distributed, times to rebound in this regime can vary widely among patients on the same therapy, even with identical values of the underlying parameters.

Survival curves, plotting the fraction of simulated patients maintaining virologic suppression over time, demonstrate the extreme interpatient variability and long follow-up times required for LRA therapy (Fig. 2C). For less than 100-fold reductions in LR size, simulated patients uniformly rebound within a few months because rebound dynamics are not in the activation-limited regime. If therapy decreases LR size 1,000-fold, then ∼55% of patients are predicted to delay rebound for at least 6 mo. However, of these patients, 47% suffer rebound in the following 6 mo. Higher reservoir reductions lead to clearance in many patients. Others, rebound may still occur after years of apparent cure, posing a challenge for patient management.

Earlier work suggested a shorter reservoir half-life of 6 mo (35), indicating that dramatic decreases in LR size would occur after 5 y or more of suppressive ART even in the absence of LRA therapy. We consider the prospects for HIV eradication or long treatment interruptions with this faster decay rate. In this optimistic scenario, only 1,500-fold reductions are needed for half of patients to clear the LR, and rebound becomes highly unlikely after a few years. Alternatively, in a worst-case scenario where latent cell death is perfectly balanced by homeostatic proliferation such that the reservoir does not decay at all (δ = 0), much higher efficacies are needed to achieve beneficial patient outcomes (Fig. 3).

### Setting Treatment Goals with Uncertainty Considerations

We conducted a full uncertainty analysis of the model, by simultaneously varying all parameters over their entire ranges (Table 1 and Fig. S1). For each simulated patient, values for the three parameters δ, A, and r were sampled independently from their respective distributions, whereas Psut was sampled from a conditional distribution that depends on r (Materials and Methods). Results for this simulated cohort are similar to those for the point estimates, with greater interpatient variation in outcomes (Fig. 3A). This variation makes the survival curves less steep: Cure is slightly more likely at low efficacy, but slightly less likely at high efficacy. As expected from Eq. 1, cure is more likely for patients with lower A or Psut values and higher δ values. If therapy provides only 10 to 100-fold LR reductions, a subset of patients may delay rebound for several months.

Using these cohort-level predictions, we can set efficacy goals for the reservoir reduction needed to achieve a particular likelihood of a desired patient outcome. Fig. 4 provides target LRA efficacies for which 50% of patients are predicted to remain rebound-free for a specified interruption time. Reductions of under 10-fold afford patients only a few weeks to a month off treatment without rebound. For 1-y interruptions, a 1,000–3,000-fold reduction is needed. To achieve the goal of eradication (cure) a 4-log reduction is required. This value increases to 4.8 logs to cure 75% of patients, and to 5.8 logs for 95% of patients.

### Model Applications and Comparison with Data

The current ability to test the model against clinical data is limited both by the dynamic range of assays measuring LR size and by the low efficacy of investigational LRA treatments. However, we can compare our predictions to results observed for non-LRA-based interventions that lead to smaller LR size and prolonged treatment interruptions (Fig. 4). A 2010 study of early ART initiators who eventually underwent treatment interruption found a single patient with LR size ∼1,500-fold lower than a typical patient (0.0064 infectious units per million resting CD4+ T cells versus an average of one per million) in whom rebound was delayed until 50 d off treatment (36). The well-known “Berlin patient” (22) has remained off treatment following a stem cell transplant since 2008, and a comprehensive analysis of his viral reservoirs found HIV DNA levels at least 7,500-fold lower than typical patients in the most sensitive assay (37). The two recently reported “Boston patients” also interrupted treatment, following transplants causing at least a 3 to 4 log decrease in viral reservoirs (23); they have since both rebounded, at ∼3 and 8 mo postinterruption. In the case of the “Mississippi baby”, infection was discovered and treated within 30 h of birth, and ART continued until interruption at around 18 mo. Virus remained undetectable for 27 mo, when viral rebound occurred, assuming the accuracy of widely reported claims (e.g., ref. 38). At the time of treatment cessation, the LR size was likely at least 300-fold lower than that of a typical adult [based on less than 0.017 infectious units per million resting CD4+ T cells at age 30 mo (39), and scaled on a weight basis relative to adults]. These few available cases demonstrate that our model is not inconsistent with experimental outcomes.
with current knowledge. When survival curves for larger cohorts become available, Bayesian methods can be used to update estimates in Table 1 and reduce uncertainty of future predictions.

Discussion

Our model is, to our knowledge, the first to quantify the required efficacy of LRAs for HIV-1 and set goals for therapy. For a wide range of parameters, we find that therapies must reduce the LR by at least 2 orders of magnitude to meaningfully increase time to rebound after ART interruption (upward infection in Figs. 2B and 3A, II; B, II; C, II), and that reductions of approximately 4 orders of magnitude are needed for half of patients to clear the infection (Figs. 3 A, I; B, I; C, I; and 4). Standard deviations in rebound times of many months are expected, owing to substantial variation in reactivation times after effective LRA therapy brings the infection to an activation-limited regime. Though the efficacy required for these beneficial outcomes likely exceeds the reach of current drugs, our results permit some optimism: We show for the first time, to our knowledge, that reactivation of all cells in the reservoir is not necessary for cessation of ART. This is because some cells in the LR will die before reactivating or, following activation, will fail to produce a chain of infections leading to rebound. On a more cautionary note, the wide distribution in reactivation times necessitates careful monitoring of patients, as rebound may occur even after long periods of viral suppression.

Even without any reservoir reduction, variation in infection parameters and chance activation together predict delays in rebound of at least 2 mo in a small minority of patients (Figs. 3 A, III and 4), consistent with ART interruption trials such as SPARTAC (40). More detailed (and possibly more speculative) models including immune responses may be needed to explain multyear posttreatment control, such as seen in the VISCONTI cohort (24).

Our analysis characterizing the required efficacy of LRA therapy does not rely on the specific mechanism of action of these drugs; only the amount by which they reduce the reservoir. We have assumed that, after ART/LRA therapy ends, cell activation and death rates return to baseline. We have also assumed that the reservoir is a homogeneous population with constant activation and death rates. The presence of reservoir compartments with different levels of LRA penetration does not alter our results, as they are stated in terms of total reservoir reduction. If, however, these compartments vary in activation or death rates (41), or if dynamics of activated cells depends on their source compartment, then our model may need to be modified. Moreover, if spatial population structure affects viral replication, viral dynamics above the detection limit (from which we estimated

Moreover, if spatial population structure affects viral replication, viral dynamics above the detection limit (from which we estimated

parameters r and A) may not correspond straightforwardly to the infection/death rates in early infection, due to local limitations in target cell density (42). Spatial restrictions on viral transmission may be particularly important in densely packed lymphoid tissue (43). Without a clear understanding of multiple compartments constituting the LR, we have considered the simplest scenario which may fit future LRA therapy outcomes.

Throughout this paper, we assume that combination ART is sufficiently effective so that viral replication alone cannot sustain the infection after all latent virus is cleared. Studies of treatment intensification (44, 45), viral evolution during ART (46, 47), and in vitro antiviral efficacy (48, 49) all support this assumption. Moreover, HIV persistence is widely believed to result solely from the long lifespan or proliferative ability of latently infected cells (3, 50). If this assumption is violated, e.g., by the presence of an altered drug-protected compartments (43, 51, 52), then any curative strategy predicated solely on latency reversal would be futile.

Our model also highlights the importance of measuring specific parameters describing latency and infection dynamics. Despite the field’s focus on measuring LR size with increasing accuracy (53), our results suggest that the rate at which latently infected cells activate, and the fraction of these that are expected to establish a rebounding infection, are more predictive of LRA outcomes. Among all parameters that determine outcome, the establishment probability is least understood, as it cannot be measured from viral load dynamics above the mutation detection limit. Simply because an integrated provirus is replication competent and transcriptionally active does not mean that it will initiate a growing infection: As with all population dynamics, chance events dominate early stages of infection growth (34, 54). HIV-1 transmission is itself a stochastic process, governed by fluctuating concentrations of early gene products (55). Sensitive assays of viral outgrowth may pave the way toward understanding the importance of these chance events to early infection; for instance, fluorescent imaging studies of adenovirus have shown that a large majority of in vitro infections seeded by single productively infected cells die out early, before rapid growth and plaque formation can occur (54). Keeping other parameters constant, assuming a worst-case (highest) value for the establishment probability raises the reservoir reductions required for cure or a desired extended rebound time by 0.8 logs. Regardless of the exact probability, the stochastic nature of HIV-1 activation and infection dynamics implies that even similarly situated patients may experience divergent responses to LRA.

The model can also advise aspects of trial design for LRAs. Survival curves computed from (58) can be used to predict the probability that a patient is cured, given that they have been off

![Fig. 3. Predicted LRA therapy outcomes, accounting for uncertainty in patient parameter values.](https://www.pnas.org/content/journals/10.1073/pnas.1406663111)

(A) Full uncertainty analysis where all viral dynamics parameters are sampled for each patient from the distributions provided in Table 1. (B) A best-case scenario where the reservoir half-life is only 6 mo (δ = 3.8 × 10⁻³ d⁻¹). All patients have the same underlying viral dynamic parameters, otherwise given by the point estimates in Table 1. (C) A worst-case scenario where the reservoir does not decay because cell death is balanced by homeostatic proliferation (δ = 0). (I) Probability that the LR is cleared by LRA. Clearance occurs if all cells in the LR die before a reactivating lineage leads to viral rebound. LRA log-efficacy is the number of orders of magnitude by which the LR size is reduced following LRA therapy, −log₁₀(q). (II) Median viral rebound times (logarithmic scale) among patients who do not clear the infection. (III) Survival curves (Kaplan–Meier plots) show the percentage of patients who have not yet experienced viral rebound, plotted as a function of the time (logarithmic scale) after treatment interruption. All simulations included 10⁵ to 10⁶ patients.
In this notation, $Y$ and $Z$ represent individual actively or latently infected cells, respectively, $\Theta$ represents no cells, and the arrows represent one type of cell becoming the other type. A latently infected cell can either activate (at rate $a$) or die (at rate $d$). An actively infected cell can either die (at rate $d$) or produce a collection of virions (at rate $b$) that results in the infection of $c$ other cells, where $c$ is a Poisson-distributed random variable with parameter $\lambda$. If $c > 0$, then $\gamma$ virions are produced (at rate $\gamma$), otherwise if $c = 0$, then $0$ virions are produced (at rate $0$). After an infection event, the original cell dies.

Each event occurs independently within a large, constant target cell population. As the model does not include limitations on viral growth, it describes only the initial stages of viral rebound. Because clinical rebound thresholds (plasma HIV RNA $>500$ copies per mL) are well below typical set points ($10^0$ to $10^1$ copies per mL), this model suffices to analyze rebound following LRA therapy and ART interruption. We do not explicitly track free virus, but assume it to be proportional to the number of infected cells. This assumption is valid because rates governing production and clearance of free virus greatly exceed other rates, allowing a separation of time scales. As we are not interested in blips or other intraday viral dynamics, this assumption does not influence our results. A method for calculating the proportionality between free virus and infected cells is provided in SI Materials and Methods.

The growth rate of the infection is $r = b + (1 - \delta) - d$. The total death rate of infected cells is $d_0 = b + d$, and the basic reproductive ratio (mean offspring for a single infected cell) is $R_0 = b/(b + d)$. The establishment probability $P_{est}$ is the solution to $P_0(t) = (1 - e^{-\delta t}) - P_{est} = 0$. The total LR decay rate in the absence of virus replication is $\delta$ if there are $Z$ cells in the LR, then the number of cells reactivating per day is $A = \delta Z$.

Analysis of the model to determine the four key parameters ($\delta$, $A$, $r$, $P_{est}$) and rapidly compute survival curves is provided in SI Materials and Methods. A script for computation of survival curves is also provided at www.danielrosenbloom.com/reboundtimes.

Parameter Estimation. The half-life of latently infected cells has been estimated to be approximately $\tau_{lat} = 44$ mo (3, 4). The resulting value of $\delta = \ln(2)/\tau_{lat}$ is centered at $5.2 \times 10^{-4}$ d$^{-1}$, and we construct a distribution of values based on ref. 3. This value represents the net rate of LR decay during suppressive therapy, considering activation, death, homeostatic proliferation, and (presumably rare) events where activated CD$^+$ T cells reenter a memory state. The net infection growth rate $\gamma$ describes the rate of exponential increase in viral load once infection has been reseeded. The LR reactivation rate $A$ is the number of cells exiting the LR per day, before LRA therapy. $A$ and $\gamma$ were jointly estimated from the dynamics of viral load during treatment interruption trials in which there was no additional reservoir-reducing intervention (5, 58); in particular, infection growth immediately following rebound is sensitive to $\gamma$, whereas the time to rebound is sensitive to $A$. Absent reservoir reduction, observed rebound dynamics are insensitive to $R_{est}$, and so this parameter was instead estimated from population genetic models (59, 60) relating observed rates of selective sweeps and emergence of drug resistance to variance in the viral offspring distribution (SI Materials and Methods).

Simulation of the Model. We use the Gillespie algorithm to track the number of latently and actively infected cells in a continuous time stochastic process. The initial number of latent cells is $Z(0) = \text{Binomial}(N_{0,LR}, Q)$, where $N_{0,LR}$ is pretreatment LR size and $Q$ is efficacy of LRA treatment (fraction of cells remaining). The initial number of actively infected cells $Y(0)$ is then chosen from a Poisson distribution with parameter $\lambda Z(0)/d_0$ (corresponding to the immigration–death equilibrium of the branching process). The simulation proceeds until the number of actively infected cells reaches the threshold for clinical detection given by a viral load of 200 copies per mL (equivalent to $3 \times 10^3$ cells total) until no active or latent cells remain. Because stochastic effects are important only for small $Y$, we switch to faster deterministic numerical integration when $Y$ reaches a level where extinction probability is very low ($< 10^{-6}$). For each $q$ value we perform $10^4$ to $10^5$ simulations.

Simulations are seeded with values of the key parameters ($\delta$, $A$, $r$, $P_{est}$), which may be either the point estimates or random numbers sampled from the distributions in Table 1. We then back out values of model-specific parameters consistent with the sampled key parameters. In general, we use a pretreatment LR size of $N_{0,LR} = 10^6$ cells to get $A = N_{0,LR}$. We then have $d_0 = \delta - a$. As detailed in SI Materials and Methods, sampling $P_{est}$ requires first
sampling the variance-to-mean ratio of the viral output distribution (ρ). Then using r and ρ along with d + c − d + b = d − 1, we can obtain i, b, d, and p0. Consistent with our generating function analysis, we find that the specific values assumed for N0 and d0 do not influence the results. For simulating other models, any other parameter assumptions are listed in the corresponding SI figure legends.

ACKNOWLEDGMENTS. We thank Y.-C. Ho, S. A. Rabi, L. Shan, and G. Laird for insightful discussions and for sharing data, and we thank A. Perelson for helpful comments on an earlier version of this manuscript. This work was supported by the Martin Delaney Collaboratory of AIDS Researchers for Eradication and the Delaney AIDS Research Enterprise Collaboratory (National Institutes of Health Grants AI096113 and U19AI090610), Grant 108165-50-RG3 from the American Foundation for AIDS Research Consortium on HIV Eradication, the Johns Hopkins Center for AIDS Research, and the Howard Hughes Medical Institute. M.A.N. was supported by the John Templeton Foundation, F.F. was funded by European Research Council Advanced Grant PDR268540, and A.L.H. and D.I.S.R. were supported by Bill and Melinda Gates Foundation Grand Challenges Explorations Grant OPP1044503.