

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

Alison L. Hill<sup>1,2,\*</sup>, Daniel I. S. Rosenbloom<sup>1,3,\*</sup>, Feng Fu<sup>4</sup>, Martin A. Nowak<sup>1</sup>,  
and Robert F. Siliciano<sup>5</sup>

<sup>1</sup>Program for Evolutionary Dynamics, Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

<sup>2</sup>Biophysics Program and Harvard-MIT Division of Health Sciences and Technology, Harvard University, Cambridge, MA 02138, USA

<sup>3</sup>Department of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032, USA

<sup>4</sup>Institute of Integrative Biology, ETH Zurich, 8092 Zurich, Switzerland

<sup>5</sup>Department of Medicine, Johns Hopkins University School of Medicine, and Howard Hughes Medical Institute, Baltimore, MD 21205, USA

\*These authors contributed equally to the manuscript

March 18, 2014

## Abstract

Massive research efforts are now underway to develop a cure for HIV infection, allowing patients to discontinue lifelong combination antiretroviral therapy (ART). New latency-reversing agents (LRAs) may be able to purge the persistent reservoir of latent virus in resting memory CD4<sup>+</sup> T cells, but the degree of reservoir reduction needed for cure remains unknown. Here we use a stochastic model of infection dynamics to estimate the efficacy of LRA needed to prevent viral rebound after ART interruption. We incorporate clinical data to estimate population-level parameter distributions and outcomes. Our findings suggest that approximately 2,000-fold reductions are required to permit a majority of patients to interrupt ART for one year without rebound and that rebound may occur suddenly after multiple years. Greater than 10,000-fold reductions may be required to prevent rebound altogether. Our results predict large variation in rebound times following LRA therapy, which will complicate clinical management. This model provides benchmarks for moving LRAs from the lab to the clinic and can aid in the design and interpretation of clinical trials. These results also apply to other interventions to reduce the latent reservoir and explain the observed return of viremia after months of apparent cure in recent bone marrow transplant recipients.

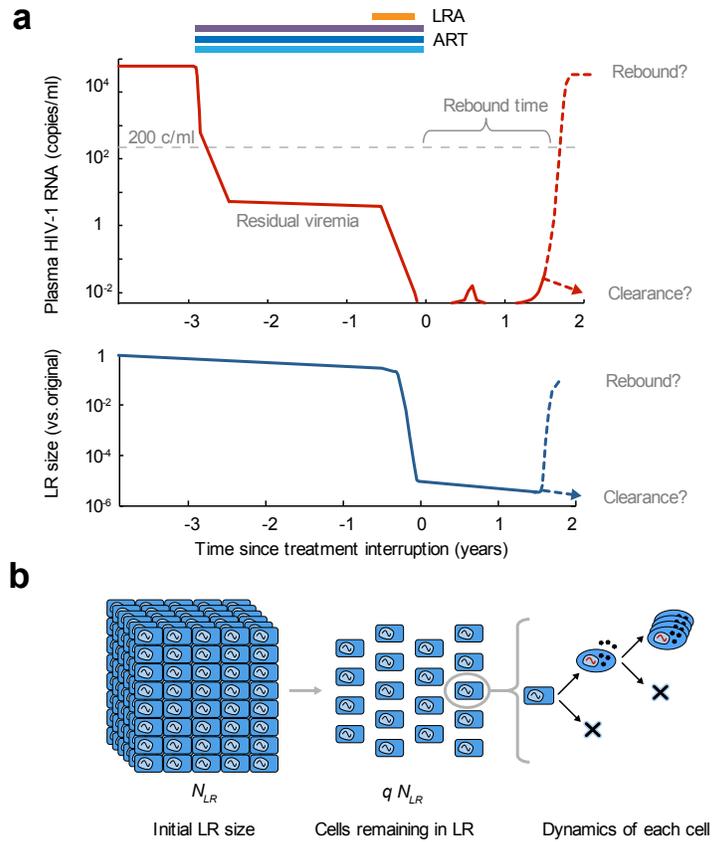
## Introduction

The latent reservoir (LR) for HIV-1 is a population of long-lived resting memory CD4<sup>+</sup> T cells with integrated HIV-1 DNA<sup>1</sup>. After establishment during acute infection<sup>2</sup>, it increases to  $10^5 - 10^7$  cells and then remains stable. As only replicating virus is targeted by antiretroviral therapy (ART), latently infected cells persist even after years of effective treatment<sup>3-7</sup>. Cellular activation leads to virus production and, if treatment is interrupted, viremia rebounds within weeks<sup>8;9</sup>. Several molecular mechanisms maintain latency, including epigenetic modifications, transcriptional interference from host genes, and the absence of activated transcription factors<sup>10-13</sup>.

Major efforts are underway to identify pharmacologic agents that reverse latency by triggering the expression of HIV-1 genes in latently infected cells, with the hope that cell death from viral cytopathic effects or cytolytic immune responses follows, reducing the size of the LR<sup>14;15</sup>. Collectively called *latency-reversing agents* (LRAs), these drugs include histone deacetylase inhibitors<sup>16-18</sup>, protein kinase C activators<sup>19-22</sup>, and the bromodomain inhibitor JQ1<sup>23-25</sup>. While LRAs are the subject of intense research, it is unclear how much the LR must be reduced to enable patients to safely discontinue ART.

The feasibility of reservoir reduction as a method of HIV-1 cure is supported by case studies of stem-cell transplantation<sup>26;27</sup> and, more recently, early treatment initiation<sup>28;29</sup>, which have allowed patients to interrupt treatment for months or years without viral rebound. 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*<sup>30;31</sup> and have not produced a measurable decrease in LR size *in vivo*<sup>16;17;32</sup>. 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 multi-phasic decay of viremia during antiretroviral therapy<sup>33;34</sup>, the initial seeding of the LR during acute infection<sup>35</sup>, the limited inflow to the LR during treatment<sup>36</sup>, the dynamics of viral blips<sup>37</sup>, and the contributions of the LR to drug resistance<sup>38</sup>. 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).



**Figure 1:** Schematic of treatment with latency-reversing agents (LRA) and stochastic model of rebound following interruption of ART. a) Proposed treatment protocol, illustrating possible viral load and size of LR before and after LRA therapy. When ART is started, viral load decreases rapidly and may fall below the limit of detection. The LR is established early in infection (not shown) and decays very slowly over time. When LRA is administered (either continuously, as shown, or in intervals), the LR declines. After discontinuation of ART, the infection may be cleared, or viremia may eventually rebound. b) LRA efficacy is defined by the parameter  $q$ , the fraction of the LR remaining after therapy, which determines the initial conditions of the model. The stochastic model of viral dynamics following interruption of ART and LRA tracks both latently infected resting  $CD4^+$  T cells (rectangles) and productively infected  $CD4^+$  T cells (ovals). Each arrow represents an event that occurs in the model. Alternate models considering homeostatic proliferation and turnover of the LR are discussed in the Methods and Supplementary Methods. Viral rebound occurs if at least one remaining cell survives long enough to activate and produce a chain of infection events leading to detectable infection (plasma HIV-1 RNA  $> 200 \text{ c ml}^{-1}$ ).

# Results

## Determination of key viral dynamic parameters governing patient outcomes

We employ a stochastic model of HIV-1 reservoir dynamics and rebound that, in its simplest form, tracks two cell types: productively infected activated CD4<sup>+</sup> T cells and latently infected resting CD4<sup>+</sup> 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 a burst of virions, resulting in the active infection of some number of other cells, or it can die from other causes without producing virions that infect other cells. The model only tracks the initial stages of viral rebound, when target cells are not yet limited. This model is similar to other stochastic models of viral dynamics<sup>39</sup>, and a full description is provided in the Methods and Supplementary Methods. The initial conditions for the dynamic model depend on the number of latently infected cells remaining following LRA therapy. LRA efficacy is defined by the fraction  $q$  of the LR that remains following therapy. 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 re-establish the infection, as it may die prior to infecting other cells. Even if it does infect others, those cells likewise may die prior to 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 protocol or mechanism of action.

We formalize the model as a multi-type branching process. Using both simulation and generating function analysis, we find that the probability and timing of rebound relies on four key parameters: the decay rate of the LR as observed during ART ( $\delta$ ), the rate at which the LR produces actively infected cells ( $A$ ), the probability that any one activated cell will produce a rebounding infection before its lineage dies ( $P_{Est}$ ), and the net growth rate of the infection once restarted ( $r$ ). Estimates of these four parameters are provided in Table 1 and Supplementary 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:

$$P_{Clr}(q) \approx e^{-qAP_{Est}/\delta}. \quad (1)$$

In the Supplementary Methods, we provide the full derivation, as well as a formula (S9) for the probability that rebound occurs a given number of days following treatment interruption (a function of  $\delta$ ,  $A$ ,  $P_{Est}$ ,  $r$ , and efficacy  $q$ ). Of note, the initial size of the reservoir itself is not included among these parameters: while it factors into both  $A$  (the product of the pre-LRA reservoir size and the per-cell activation rate), and  $q$  (the ratio of post-LRA reservoir size to pre-LRA reservoir size), it does not independently influence outcomes. Both of these formulas provide an excellent match to explicit simulation of the model (Fig. 2). The key assumption required for the analysis is that  $r$  greatly exceeds  $\delta$ ; since viral doubling times during rebound are measured on the order of a few days, while LR decay is measured on the order of many months or years, this assumption is expected to hold. Likelihood-based inference can therefore be conducted by efficient computation of

rebound probabilities (using equation (S9)), rather than by time-consuming stochastic simulation.

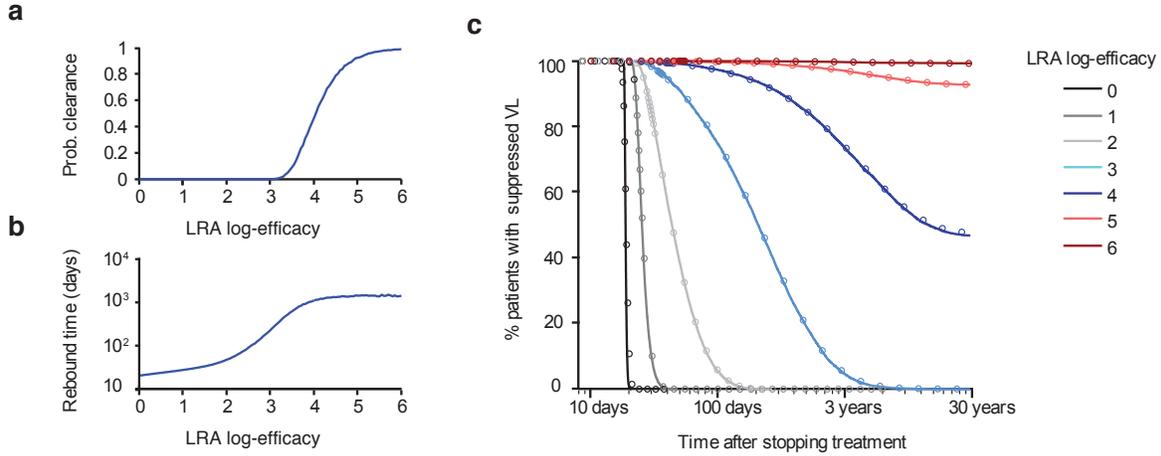
Outcomes depend only on the four parameters above even in more complex models of viral dynamics that include additional features of T cell biology and the HIV lifecycle (Supplementary Methods). Alternate models studied include explicit tracking of free virus with varying burst sizes, an “eclipse phase” during which an infected cell produces no virus, proliferation of cells upon reactivation, maintenance of the LR by homeostatic proliferation, and either a constant or Poisson-distributed number of infected cells produced by a single cell (Supplementary 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 (Supplementary Figs. S6, 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.

Parameter	Symbol	Estimation Method	Source	Best Estimate	Distribution
LR decay rate	$\delta$	Long-term ART ( $\delta = \ln(2)/\tau_{1/2}$ )	6:7	$5.2 \times 10^{-4} \text{ d}^{-1}$	$\delta \sim \mathcal{N}(5.2, 1.6) \times 10^{-4} \text{ d}^{-1}$
LR exit rate	$A$	Viral rebound after ART interruption	8:52	57 cells $\text{d}^{-1}$	$\log_{10}(A) \sim \mathcal{N}(1.76, 1.0)$ $\log_{10}(r) \sim \mathcal{N}(-0.40, 0.19)$
Growth rate	$r$			0.4 $\text{d}^{-1}$	
Establishment probability	$P_{Est}$	Population genetic modeling	53-55	0.07	(composite distribution; see Methods)

**Table 1:** Estimated values for the key parameters of the stochastic viral dynamics model. The half-life of latently infected cells has been estimated to be approximately  $\tau_{1/2} = 44$  months<sup>6:7</sup>. The resulting value of  $\delta = \ln(2)/\tau_{1/2}$  is centered at  $5.2 \times 10^{-4} \text{ d}^{-1}$ , and we construct a distribution of values based on the earlier study. This value represents the *net* rate of LR decay during suppressive therapy, considering activation, death, homeostatic proliferation, and (presumably rare) events where activated CD4<sup>+</sup> T cells re-enter a memory state. The net infection growth rate  $r$  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 reservoir-reducing therapy.  $A$  and  $r$  were jointly estimated from the dynamics of viral load during treatment interruption<sup>8:52</sup>; in particular, infection growth immediately following rebound is sensitive to  $r$ , while the time to rebound is sensitive to  $A$ . Finally, the establishment probability  $P_{Est}$  was estimated using population genetic models<sup>53-55</sup> that relate observed rates of selective sweeps and emergence of drug resistance to variance in the viral offspring distribution (see Methods). Notation  $X \sim \mathcal{N}(\mu, \sigma)$  means that  $X$  is a random variable drawn from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

## Predicted prospects for eradicating infection or delaying time to rebound

Using best estimates of parameters (Table 1), we can explore the likely outcomes of interventions that reduce the latent reservoir. 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



**Figure 2:** Clearance probabilities and rebound times following LRA therapy predicted from model using point estimates for the parameters (Table 1). “LRA log-efficacy” is the number of orders of magnitude by which the latent reservoir size is reduced following LRA therapy ( $-\log_{10}(q)$ ). a) 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. b) Median viral rebound times (logarithmic scale), among patients who do not clear the infection. c) 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. Solid lines represent simulations, and circles represent approximations from the branching process calculation. All simulations included  $10^4$  to  $10^5$  patients with identical viral dynamic parameter values.

infection during the patient’s lifespan. In this case, LRA has essentially cleared the infection and a cure is achieved. We simulated the model to predict the relationship between LRA efficacy and clearance (Fig. 2a). We find that the reservoir 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 outcome is extension of the time until rebound, defined as plasma HIV-1 RNA  $\geq 200$  c ml<sup>-1</sup>. 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 the reservoir size is needed for median rebound times of 1 year. Only modest ( $\sim 2$ -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  $\approx 100$ -fold reduction and eventually reaches a ceiling as 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*. Since waiting time is roughly exponentially

distributed, times to rebound in this regime can vary widely among patients experiencing the same therapy, even with identical values of the underlying parameters.

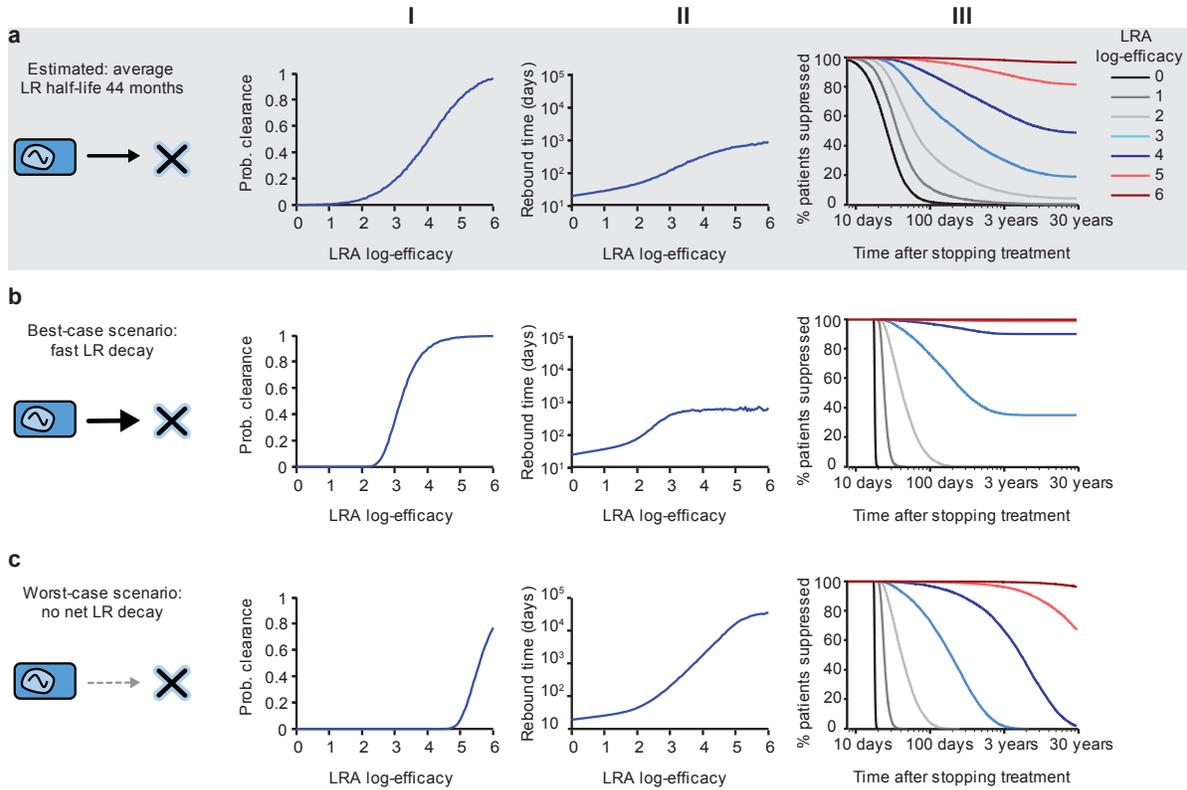
Survival curves, plotting the fraction of simulated patients maintaining virologic suppression over time, demonstrate the extreme interpatient variability and long follow-up times required for LRA therapy (Fig. 2c). For less than 100-fold reductions in LR size, simulated patients uniformly rebound within a few months, since rebound dynamics are not in the activation-limited regime. If therapy decreases LR size 1,000-fold, then  $\sim 55\%$  of patients are predicted to delay rebound for at least six months. However, of these patients, 47% suffer rebound in the following six months. Higher reservoir reductions lead to clearance in many patients. In 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 months<sup>40</sup>, indicating that dramatic decreases in LR size would occur after 5 or more years of suppressive ART even in the absence of LRA therapy. We consider the prospects for HIV eradication or long treatment interruptions with this faster reservoir decay rate (Fig. 3b). 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 ( $\delta = 0$ ), much higher efficacies are needed to achieve beneficial patient outcomes (Fig. 3c).

## 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, Supplementary Fig. S1). For each simulated patient, values for the three parameters  $\delta$ ,  $A$ , and  $r$  were sampled independently from their respective distributions, while  $P_{Est}$  was sampled from a conditional distribution that depends on  $r$  (see 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 Equation (1), cure is more likely for patients with lower  $A$  or  $P_{Est}$  values and higher  $\delta$  values. If therapy provides only 10–100-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 the target LRA efficacies for which 50% of patients are predicted to remain rebound-free for a specified interruption time. Reductions of less than 10-fold afford patients only a few weeks to a month off treatment without rebound. For one-year interruptions, a 1,000–3,000-fold reduction is required. To achieve the goal of eradication (or cure) a 4-log reduction is required. This value increases to 4.8 logs if we desire eradication in 75% of patients, and to 5.8 logs to cure 95% of patients.



**Figure 3:** Predicted LRA therapy outcomes, accounting for uncertainty in patient parameter values. 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 months ( $\delta = 3.8 \times 10^{-3}$ ). 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 ( $\delta = 0$ ). 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 latent reservoir size is reduced following LRA therapy ( $-\log_{10}(q)$ ). 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  $10^4$  to  $10^5$  patients.

## Model applications and comparison to data

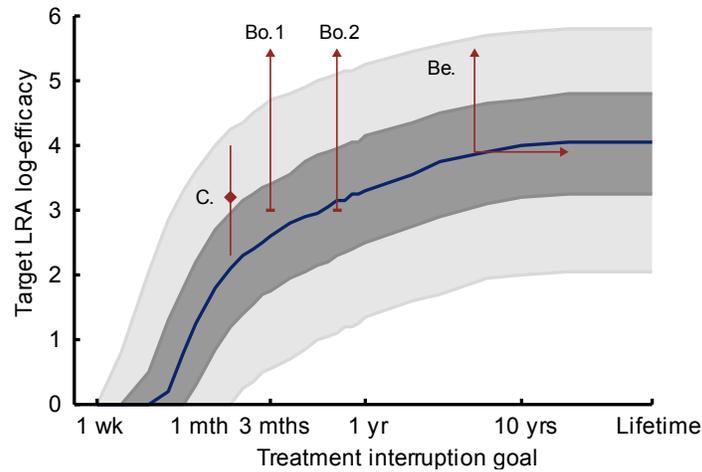
Current ability to test the model against clinical data is limited both by the dynamic range of assays measuring LR size and by the low efficacy of investigational LRA treatments. Yet we can compare our predictions to results observed for non-LRA-based interventions that lead to decreases in LR size 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 approximately 1,500-fold lower than a typical patient (0.0064 infectious units per million resting CD4<sup>+</sup> T cells, versus an average of 1 per million) in whom rebound was delayed until 50 days off treatment<sup>41</sup>. The well-known ‘Berlin patient’<sup>26</sup> 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<sup>42</sup>. The two recently reported ‘Boston patients’ also interrupted treatment, following transplants that resulted in “at least a 3 to 4 log<sub>10</sub> decrease” in viral reservoirs<sup>27</sup>; they have since both rebounded, at approximately 3 and 8 months post-interruption. These few available cases demonstrate that our model is not inconsistent with current knowledge. When survival curves for larger cohorts become available, Bayesian methods can be used to update the estimates in Table 1 and reduce uncertainty of future predictions.

## Discussion

Our model is the first to quantify the required efficacy of latency-reversing agents 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 two orders of magnitude to produce a meaningful increase in the time to rebound after ART interruption (upward inflection in Figs. 2b, 3II), and that reductions of approximately four orders of magnitude are needed for half of patients to clear the infection (Figs. 3a, 4). Standard deviations in rebound times of many months are expected, owing to variation in pretreatment reservoir size and exponentially-distributed reactivation times after effective LRA therapy brings the infection to an activation-limited regime. While the efficacy required for these beneficial outcomes is likely beyond the reach of current drugs, our results permit some optimism: we show for the first time 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 infection events leading to rebound. On a more cautionary note, the wide distribution in reactivation times necessitates careful monitoring of patients, as rebound is possible even after long periods of viral suppression.

Even without any reservoir reduction, variation in infection parameters and stochastic activation together predict delays in rebound of at least two months in a small minority of patients (Fig. 4), consistent with ART interruption trials such as SPARTAC<sup>43</sup>. More detailed (and possibly more speculative) models including specific immune responses may be needed to explain multi-year post-treatment control, such as found in the VISCONTI cohort<sup>28</sup>.

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



**Figure 4:** Efficacies required for successful LRA therapy. The target LRA log-efficacy is the treatment level (in terms of log-reduction in latent reservoir size) for which at least 50% of patients still have suppressed viral load after a given treatment interruption length (blue line). Shaded ranges show the results for the middle 50% (dark gray) and 90% (light gray) of patients. “Lifetime” means the LR is cleared. Annotations on the curve represent data points for case studies describing large reservoir reductions and observing rebound times after ART interruption. From left to right, they represent a case of early ART initiation (the “Chun patient”<sup>41</sup>, diamond), and three cases of hematopoietic stem cell transplant: the two “Boston patients”<sup>27</sup> (vertical arrows) and the “Berlin patient”<sup>26;42</sup> (vertical and horizontal arrow). For the Chun patient, the annotations represent the maximum likelihood estimate for LR reduction (diamond), as well as 95% confidence intervals (vertical bar). For the Boston and Berlin patients, vertical arrows indicate that only a lower bound on treatment efficacy is known (LR size was below the detection limit) and that the true value may extend further in the direction shown. For the Berlin patient, the horizontal arrow indicates that rebound time is at least five years (rebound has not yet occurred).

have also assumed that the reservoir is a homogeneous population without variation in activation and death rates. The presence of reservoir compartments with different levels of drug 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<sup>44</sup>, or if viral dynamics of activated cells depends on their source compartment, then our model may need to be modified. In the absence of clear understanding of multiple compartments constituting the LR, we have considered the simplest scenario which may be able to fit future LRA therapy outcomes.

Our model also highlights the importance of measuring specific parameters describing latency and infection dynamics. Despite the field's focus on measuring latent reservoir size with increasing accuracy<sup>45;46</sup>, 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 limit of detection. 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<sup>39;47</sup>. Full-length HIV-1 transcription is itself a stochastic process, governed by fluctuating concentrations of early gene products<sup>48</sup>. Sensitive assays of viral outgrowth may pave the way toward understanding the importance of these chance events to early infection; for example, ongoing experiments using fluorescent imaging of *in vitro* infections seeded by single cells productively infected with a reporter HIV strain suggest that only 14% trigger a growing infection before their lineage dies out (G. Laird, personal communication). The stochastic nature of HIV-1 infection dynamics implies that even similarly situated patients may experience divergent responses to LRA therapy.

The model can also advise aspects of trial design for LRAs. Survival curves computed from equation (S9) can be used to predict the probability that a patient is cured, given that they have been off treatment without rebound for a known period. As frequent viral load testing for years of post-interruption monitoring is not feasible, it may be helpful to choose sampling timepoints based on the expected distribution of rebound times. Trial design is complicated by the fact that LRA treatment efficacy is unknown if post-treatment LR size is below the detection limit. By considering prior knowledge about viral dynamics parameters and the range of possible treatment efficacies, the model can be used to estimate outcomes even in the presence of uncertainty.

To date, laboratory and clinical studies of investigational LRAs have generally found weak potential for reservoir reduction — up to one log-reduction *in vitro* and less *in vivo*<sup>16;30;49</sup>. We predict that much higher efficacy will be required for eradication, which may be achieved by multiple rounds of LRA therapy, a combination of therapies, or development of therapies to which a greater fraction of the LR is susceptible. While we have focused on LRA therapy, our findings also serve to interpret infection eradication or delays in rebound caused by early treatment<sup>28;29;50;51</sup> or stem cell transplantation<sup>26;27</sup>, both of which also reduce the latent reservoir. We believe that these modeling efforts will provide a quantitative framework for interpreting clinical trials of any reservoir-reduction strategy.

## Acknowledgements

We thank Ya-chi Ho, S. Alireza Rabi, Liang Shan, and Greg Laird for insightful discussions and for sharing data. This work was supported by the Martin Delaney CARE and DARE Collaboratories (NIH grants AI096113 and 1U19AI096109), by an ARCHE Collaborative Research Grant from the Foundation for AIDS Research (amFAR 108165-50-RGRL), by the Johns Hopkins Center for AIDS Research, by NIH grant 43222, and by the Howard Hughes Medical Institute. MAN was supported by the John Templeton Foundation. FF was funded by a European Research Council Advanced Grant (PBDR 268540). ALH and DISR were supported by a Bill & Melinda Gates Foundation Grand Challenges Explorations Grant (OPP1044503).

## References

- [1] Chun, T. W., *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**(6629), 183–188 (1997).
- [2] Chun, T. W., *et al.* Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection. *Proc. Natl. Acad. Sci. USA* **95**(15), 8869–8873 (1998).
- [3] Finzi, D., *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**(5341), 1295–1300 (1997).
- [4] Wong, J. K., *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**(5341), 1291–1295 (1997).
- [5] Chun, T. W., *et al.* Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl. Acad. Sci. USA* **94**(24), 13193–13197 (1997).
- [6] Siliciano, J. D., *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat. Med.* **9**(6), 727–728 (2003).
- [7] Archin, N. M., *et al.* Measuring HIV latency over time: Reservoir stability and assessing interventions. In *21st Conference on Retroviruses and Opportunistic Infections*, Boston, MA, (2014).
- [8] Ruiz, L., *et al.* Structured treatment interruption in chronically HIV-1 infected patients after long-term viral suppression. *AIDS* **14**(4), 397 (2000).
- [9] Davey, Jr., R. T., *et al.* HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. *Proc. Natl. Acad. Sci. USA* **96**(26), 15109–15114 (1999).
- [10] Marsden, M. D. & Zack, J. A. Establishment and maintenance of HIV latency: model systems and opportunities for intervention. *Future Virol* **5**(1), 97–109 (2010).

- [11] Hakre, S., Chavez, L., Shirakawa, K., & Verdin, E. Epigenetic regulation of HIV latency. *Curr. Opin. HIV AIDS* **6**(1), 19–24 (2011).
- [12] Mbonye, U. & Karn, J. Control of HIV latency by epigenetic and non-epigenetic mechanisms. *Curr HIV Res.* **9**(8), 554–567 (2011).
- [13] Ruelas, D. S. & Greene, W. C. An integrated overview of HIV-1 latency. *Cell* **155**(3), 519–529 (2013).
- [14] Choudhary, S. K. & Margolis, D. M. Curing HIV: pharmacologic approaches to target HIV-1 latency. *Ann. Rev. Pharmacol. Toxicol.* **51**(1), 397–418 (2011).
- [15] Durand, C. M., Blankson, J. N., & Siliciano, R. F. Developing strategies for HIV-1 eradication. *Trends Immunol* **33**(11), 554–562 (2012).
- [16] Archin, N. M., *et al.* Antiretroviral intensification and valproic acid lack sustained effect on residual HIV-1 viremia or resting CD4+ cell infection. *PLoS ONE* **5**(2), e9390 (2010).
- [17] Archin, N. M., *et al.* Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487**(7408), 482–485 (2012).
- [18] Shirakawa, K., Chavez, L., Hakre, S., Calvanese, V., & Verdin, E. Reactivation of latent HIV by histone deacetylase inhibitors. *Trends Microbiol* **21**(6), 277–285 (2013).
- [19] Korin, Y. D., Brooks, D. G., Brown, S., Korotzer, A., & Zack, J. A. Effects of prostratin on T-cell activation and human immunodeficiency virus latency. *J Virol* **76**(16), 8118–8123 (2002).
- [20] Williams, S. A., *et al.* Prostratin antagonizes HIV latency by activating NF-kappaB. *J. Biol Chem* **279**(40), 42008–42017 (2004).
- [21] Mehla, R., *et al.* Bryostatin modulates latent HIV-1 infection via PKC and AMPK signaling but inhibits acute infection in a receptor independent manner. *PloS ONE* **5**(6), e11160 (2010).
- [22] DeChristopher, B. A., *et al.* Designed, synthetically accessible bryostatin analogues potently induce activation of latent HIV reservoirs in vitro. *Nat Chem* **4**(9), 705–710 (2012).
- [23] Bartholomeeusen, K., Xiang, Y., Fujinaga, K., & Peterlin, B. M. Bromodomain and extra-terminal (BET) bromodomain inhibition activate transcription via transient release of positive transcription elongation factor b (p-TEFb) from 7SK small nuclear ribonucleoprotein. *J. Biol Chem* **287**(43), 36609–36616 (2012).
- [24] Zhu, J., *et al.* Reactivation of latent HIV-1 by inhibition of BRD4. *Cell Rep.* **2**(4), 807–816 (2012).
- [25] Boehm, D., *et al.* BET bromodomain-targeting compounds reactivate HIV from latency via a tat-independent mechanism. *Cell Cycle* **12**(3), 452–462 (2013).

- [26] Hütter, G., *et al.* Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *New Engl J Med* **360**(7), 692–698 (2009). PMID: 19213682.
- [27] Henrich, T. J., *et al.* HIV-1 rebound following allogeneic stem cell transplantation and treatment interruption. In *21st Conference on Retroviruses and Opportunistic Infections*, Boston, MA, (2014).
- [28] Sáez-Cirión, A., *et al.* Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI study. *PLoS Pathog* **9**(3), e1003211 (2013). PMID: 23516360.
- [29] Persaud, D., *et al.* Absence of detectable HIV-1 viremia after treatment cessation in an infant. *New Engl J Med* **369**(19), 1828–1835 (2013). PMID: 24152233.
- [30] Cillo, A. Only a small fraction of HIV-1 proviruses in resting CD4+ T cells can be induced to produce virions *ex vivo* with anti-CD3/CD28 or vorinostat. In *20th Conference on Retroviruses and Opportunistic Infections*, Atlanta, GA, (2013).
- [31] Bullen, C. K., Laird, G. M., Durand, C. M., Siliciano, J. D., & Siliciano, R. F. New *ex vivo* approaches distinguish effective and ineffective single agents for reversing HIV-1 latency *in vivo*. *Nat. Med.* **in press**.
- [32] Spivak, A. M., *et al.* A pilot study assessing the safety and latency-reversing activity of disulfiram in HIV-1 infected adults on antiretroviral therapy. *Clin Infect Dis* **58**(6), 883–890 (2014). PMID: 24336828.
- [33] Perelson, A. S., *et al.* Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature* **387**(6629), 188–191 (1997).
- [34] Nowak, M. A. & May, R. M. C. *Virus Dynamics: Mathematical principles of immunology and virology*. Oxford Univ. Press, USA, (2000).
- [35] Archin, N. M., *et al.* Immediate antiviral therapy appears to restrict resting CD4+ cell HIV-1 infection without accelerating the decay of latent infection. *Proc. Natl. Acad. Sci. USA* **109**(24), 9523–9528 (2012).
- [36] Sedaghat, A. R., Siliciano, J. D., Brennan, T. P., Wilke, C. O., & Siliciano, R. F. Limits on replenishment of the resting CD4+ T cell reservoir for HIV in patients on HAART. *PLoS Pathog* **3**(8), e122 (2007).
- [37] Conway, J. M. & Coombs, D. A stochastic model of latently infected cell reactivation and viral blip generation in treated HIV patients. *PLoS Comput Biol* **7**(4), e1002033 (2011).
- [38] Rosenbloom, D. I. S., Hill, A. L., Rabi, S. A., Siliciano, R. F., & Nowak, M. A. Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. *Nat. Med.* **18**(9), 1378–1385 (2012).

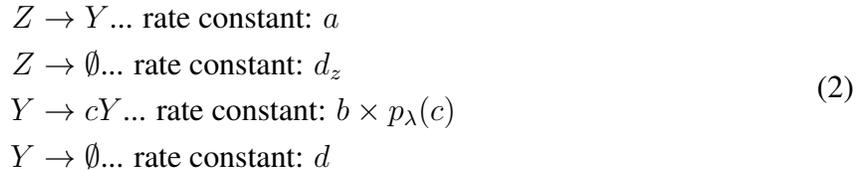
- [39] Pearson, J. E., Krapivsky, P., & Perelson, A. S. Stochastic theory of early viral infection: Continuous versus burst production of virions. *PLoS Comput Biol* **7**(2), e1001058 (2011).
- [40] Zhang, L., *et al.* Quantifying residual HIV-1 replication in patients receiving combination antiretroviral therapy. *New Engl J Med* **340**(21), 1605–1613 (1999). PMID: 10341272.
- [41] Chun, T.-W., *et al.* Rebound of plasma viremia following cessation of antiretroviral therapy despite profoundly low levels of HIV reservoir: implications for eradication. *AIDS* **24**(18), 2803–2808 (2010). PMID: 20962613 PMCID: PMC3154092.
- [42] Yukl, S. A., *et al.* Challenges in detecting HIV persistence during potentially curative interventions: A study of the Berlin Patient. *PLoS Pathog* **9**(5), e1003347 (2013).
- [43] Stöhr, W., *et al.* Duration of HIV-1 viral suppression on cessation of antiretroviral therapy in primary infection correlates with time on therapy. *PloS one* **8**(10), e78287 (2013).
- [44] Buzón, M. J., *et al.* HIV-1 persistence in CD4(+) T cells with stem cell-like properties. *Nat Med* **20**(2), 139–142 (2014).
- [45] Laird, G. M., *et al.* Rapid quantification of the latent reservoir for HIV-1 using a viral outgrowth assay. *PLoS Pathog* **9**(5), e1003398 (2013). PMID: 23737751 PMCID: PMC3667757.
- [46] Ho, Y.-C., *et al.* Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**(3), 540–551 (2013). PMID: 24243014.
- [47] Hofacre, A., Wodarz, D., Komarova, N. L., & Fan, H. Early infection and spread of a conditionally replicating adenovirus under conditions of plaque formation. *Virology* **423**(1), 89–96 (2012). PMID: 22192628.
- [48] Singh, A. & Weinberger, L. S. Stochastic gene expression as a molecular switch for viral latency. *Curr Opin in Microbiol* **12**(4), 460–466 (2009).
- [49] Xing, S., *et al.* Disulfiram reactivates latent HIV-1 in a bcl-2-transduced primary CD4+ T cell model without inducing global T cell activation. *J. Virol.* **85**(12), 6060–6064 (2011).
- [50] Blankson, J. N., Siliciano, J. D., & Siliciano, R. F. The Effect of early treatment on the latent reservoir of HIV-1. *J. Infect. Dis.* **191**(9), 1394–1396 (2005).
- [51] Strain, M. C., *et al.* Effect of treatment, during primary infection, on establishment and clearance of cellular reservoirs of HIV-1. *J. Infect. Dis.* **191**(9), 1410–1418 (2005).
- [52] Luo, R., Piovoso, M. J., Martinez-Picado, J., & Zurakowski, R. HIV model parameter estimates from interruption trial data including drug efficacy and reservoir dynamics. *PLoS ONE* **7**(7), e40198 (2012).
- [53] Rouzine, I. M. & Coffin, J. M. Linkage disequilibrium test implies a large effective population number for HIV in vivo. *Proc. Natl. Acad. Sci. USA* **96**(19), 10758–10763 (1999).

- [54] Pennings, P. S. Standing genetic variation and the evolution of drug resistance in HIV. *PLoS Comput Biol* **8**(6), e1002527 (2012).
- [55] Pennings, P. S., Kryazhimskiy, S., & Wakeley, J. Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genetics* **10**(1), e1004000 (2014).
- [56] Eriksson, S., *et al.* Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies. *PLoS Pathog* **9**(2), e1003174 (2013).
- [57] Karlin, S. & Taylor, H. E. *A First Course in Stochastic Processes*. Academic Press, San Diego, (1975).
- [58] De Boer, R. J., Ribeiro, R. M., & Perelson, A. S. Current estimates for HIV-1 production imply rapid viral clearance in lymphoid tissues. *PLoS Comput Biol* **6**(9), e1000906 (2010).
- [59] Di Mascio, M., *et al.* Noninvasive in vivo imaging of CD4 cells in simian-human immunodeficiency virus (SHIV)-infected nonhuman primates. *Blood* **114**(2), 328–337 (2009).
- [60] Ganusov, V. V. & De Boer, R. J. Do most lymphocytes in humans really reside in the gut? *Trends Immunol* **28**(12), 514–518 (2007).
- [61] Chen, H. Y., Di Mascio, M., Perelson, A. S., Ho, D. D., & Zhang, L. Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. *Proc. Natl. Acad. Sci. USA* **104**(48), 19079–19084 (2007).
- [62] Reilly, C., Wietgreffe, S., Sedgewick, G., & Haase, A. Determination of simian immunodeficiency virus production by infected activated and resting cells. *AIDS* **21**(2), 163–168 (2007).
- [63] Markowitz, M., *et al.* A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *J. Virol.* **77**(8), 5037–5038 (2003).
- [64] Ribeiro, R. M., *et al.* Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. *Journal of Virology* **84**(12), 6096–6102 (2010).
- [65] Sigal, A., *et al.* Cell-to-cell spread of HIV permits ongoing replication despite antiretroviral therapy. *Nature* **477**(7362), 95–98 (2011).
- [66] Kouyos, R. D., Althaus, C. L., & Bonhoeffer, S. Stochastic or deterministic: what is the effective population size of HIV-1? *Trends Microbiol* **14**(12), 507–511 (2006).
- [67] Abate, J., Choudhury, G. L., & Whitt, W. An introduction to numerical transform inversion and its application to probability models. In *Computational Probability*, Grassmann, W. K., editor, number 24 in *International Series in Operations Research & Management Science*, 257–323. Springer US (2000).

# Methods

## Stochastic model of viral dynamics

The basic model of reservoir dynamics and rebound tracks two cell types: productively infected activated CD4<sup>+</sup> T cells, and latently infected resting CD4<sup>+</sup> T cells. Cells with nonviable provirus, which may vastly outnumber those with replication-competent provirus<sup>56</sup>, are excluded. The model can be described formally as a two-type branching process, in which four types of events can occur (Fig. 1):



In this notation  $Y$  and  $Z$  represent individual actively or latently infected cells, respectively,  $\emptyset$  represents no cells, and the arrows represent events that lead one type of cell to become the other type. A latently infected cell can either activate (at rate  $a$ ) or die (at rate  $d_z$ ). An actively infected cell can either die (at rate  $d$ ) or produce a burst of virions (at rate  $b$ ) that results in the infection of  $c$  other cells, where  $c$  is a Poisson-distributed random variable with parameter  $\lambda$ ,  $p_\lambda(c) = (\exp(-\lambda)\lambda^c)/(c!)$ . After a burst event, the original cell dies. This model is similar to other stochastic models of viral dynamics<sup>39</sup>.

Each infection is independent and occurs 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. Since clinical rebound thresholds (plasma HIV RNA  $> 50 - 200$  c ml<sup>-1</sup>) are well below typical setpoints ( $10^4 - 10^6$  c ml<sup>-1</sup>), this model suffices to analyze rebound following LRA therapy and ART interruption. We do not explicitly track free virus, but assume it is at a level proportional to the number of infected cells. This assumption is valid because the rates governing the production of virus from infected cells and the clearance rate of free virus are much higher than other rates, allowing a separation of time scales. Because 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 below.

The growth rate of the infection is  $r = b(\lambda - 1) - d$ . The total death rate of infected cells is  $d_y = b + d$ , and the basic reproductive ratio (mean offspring number for a single infected cell) is  $R_0 = \frac{b\lambda}{b+d}$ . The establishment probability  $P_{Est}$  is the solution to  $R_0(1 - e^{-\lambda P_{Est}}) - \lambda P_{Est} = 0$ .

## Generating function analysis of model

A simplified version of the above branching process was used, for which probability generating functions could be written and expanded explicitly. In this simplified process, each birth event results in exactly two actively infected cells. The initial number of latently infected cells was treated as Poisson-distributed around expected value  $qN_{LR}$ , where  $N_{LR}$  is the pre-therapy LR size and  $q$

represents therapy efficacy. The initial number of actively infected cells was treated as Poisson-distributed around  $aqN_{LR}/(b+d)$ , representing activation-death equilibrium during effective treatment (birth events failing to produce new infected cells) with latently infected cells held constant at  $qN_{LR}$ . Latent cell dynamics following interruption were the same as in the model described above. The backwards Kolmogorov equations for this process<sup>57</sup> can be solved explicitly, resulting in a probability generating function for the number of actively infected cells existing at time  $t$ . In the Supplementary Methods, we derive the clearance probability  $P_{Clr} \approx \exp[-qAP_{Est}/\delta]$  and the probability that there are  $\mathcal{Y}$  actively infected cells at time  $t$ :

$$\begin{aligned}
P(\mathcal{Y}, t) \approx & \exp \left[ -qAP_{Est} \left( \omega_1 + \frac{1 - e^{\delta t} \left( 1 + \delta \int_0^t g(\tau) d\tau \right)}{\delta P_{Est}} \right) \right] \\
& \times \frac{1}{\mathcal{Y}!} \left( \frac{e^{rt} - 1}{e^{rt} - 1 + P_{Est}} \right)^{\mathcal{Y}} \\
& \times (\omega_2)_{\mathcal{Y}} M \left( -\mathcal{Y}, \omega_2, -\frac{qAP_{Est}^2}{2 - P_{Est}} \omega_1 \right),
\end{aligned} \tag{3}$$

where  $\omega_1 = (e^{rt} P_{Est}) / (r(e^{rt} - 1)(e^{rt} - 1 + P_{Est}))$ ,  $\omega_2 = (qAP_{Est}e^{-\delta t}) / r$ ,  $g(t) = e^{\delta t} (e^{rt} - 1)(1 - P_{Est}) / (e^{rt} - 1 + P_{Est})$ ,  $(\cdot)_{\mathcal{Y}}$  is the Pochhammer symbol (rising factorial), and  $M$  is Kummer's confluent hypergeometric function. Both approximations hold for  $\delta$  much smaller than  $r$ . In the Supplementary Methods, we also provide an efficient method for calculating rebound probabilities without repeated evaluation of Eq. (3) at multiple values of  $\mathcal{Y}$ . This efficient method allows for computation of all survival curves in Fig. 2 in seconds. A script is provided at <http://www.danielrosenbloom.com/reboundtimes> to rapidly compute survival curves using these methods.

## Estimating the cell-to-virus ratio

Because the model tracks only infected cells, but must relate total body cell counts to clinically observed plasma viral RNA concentrations, we require an estimate of the conversion factor between these two quantities. For this calculation we applied a version of the analysis developed in De Boer *et al.*<sup>58</sup>. Let  $\mathcal{Y}$  be the total body number of productively infected cells residing in lymphoid tissue, let  $\mathcal{V}$  be the total body number of free virus particles, and let  $\kappa$  be the per-cell production rate of virions. We can balance viral production and decay at equilibrium for a typical 70-kg individual:

$$\begin{aligned}
\kappa \mathcal{Y} &= d_v \mathcal{V} \\
&= \left( \frac{d_v}{\kappa} \right) \left( \frac{100 \text{ virions in LT}}{1 \text{ virion in ECF}} \right) \left( \frac{1 \text{ virion in ECF}}{2 \text{ RNA copies in ECF}} \right) (15,000 \text{ ml ECF}) v \\
&= \left( \frac{d_v}{\kappa} \right) \left( 7.5 \times 10^5 \frac{\text{virions in LT}}{\text{RNA copies per ml ECF}} \right) v,
\end{aligned} \tag{4}$$

where  $v$  is the per-ml concentration of viral RNA in circulation. Here we use the estimate that virus particles in the lymphoid tissue outnumber those in circulation 100-fold, mirroring the ratio

of lymphocytes in lymph nodes versus in circulation<sup>59;60</sup>. Note that since the lifespan of an infected cell is much longer than that of a free virion, the ratio of production to decay rates,  $\kappa/d_v$ , equals the number of free virions associated with a productive cell during most of that cell’s existence; values of 200–1000 are consistent with recent experiments<sup>58;61;62</sup>. The resulting estimate for  $\mathcal{Y}$  is therefore  $(750 \text{ to } 3750) \times v$ . We use the geometric mean of these values – 1680 – as a point estimate. With this value the viral rebound threshold of 200 c ml<sup>-1</sup> corresponds to  $\mathcal{Y} \approx 3 \times 10^5$ . Note that model results are insensitive to this value, as rebound probability depends on the logarithm of the rebound threshold.

## Estimating key parameters

### Estimation of $\delta$

The net decay rate of the latent reservoir ( $\delta$ ) was estimated from a previous study of longitudinal reservoir size in patients on long-term ART. Table 1 of Siliciano *et al.*<sup>6</sup> reports the mean and 95% confidence interval of LR decay from a cohort of 59 patients, derived using a mixed-effects model. This method models each patient’s decay rate  $\delta_i$  as being sampled from a population-level normal distribution with mean  $5.2 \times 10^{-4} \text{ d}^{-1}$  and standard deviation of  $1.6 \times 10^{-4}$ . We use this distribution to sample simulated patients. Note that with this method, approximately 1 in 1500 draws will give a negative value of  $\delta$ , corresponding to an LR that fails to decay (e.g., by homeostatic proliferation outweighing activation plus death). We allowed negative  $\delta$  only for simulation of the model with homeostatic proliferation; for all others, we truncated the distribution at zero.

The mean decay rate from the Siliciano study corresponds to a reservoir half-life of 44 months, which is consistent with a more recent study that found a value of 43 months<sup>7</sup>.

### Estimation of $r$

The net exponential growth rate of the infection ( $r$ ) and the number of latent cells reactivating per day ( $A$ ) before reservoir-reducing therapy can be estimated from ART-interruption studies. Luo *et al.*<sup>52</sup> report the joint posterior distributions of viral dynamic parameters for 10 patients who underwent 3-5 structured treatment interruptions in a previous study<sup>8</sup>. From these, we computed inter-patient distributions for both  $r$  and  $A$ .

The authors used a system of three differential equations to describe viral rebound during treatment interruption:  $\dot{x}(t) = \lambda_x - d_x x(t) - \beta x(t)v(t)$ ,  $\dot{y}(t) = \beta x(t)v(t) - d_y y(t) + \lambda_y$ , and  $\dot{v}(t) = \gamma y(t) - d_v v(t)$ ; where state variables  $x$ ,  $y$ , and  $v$  represented plasma concentrations of uninfected CD4<sup>+</sup> T cells, productively infected CD4<sup>+</sup> T cells, and free virus, respectively. Parameters  $\lambda_x$  and  $d_x$  denote production and death rates of target cells, respectively;  $\beta$  is infectivity;  $d_y$  is the total death rate of infected cells;  $\gamma$  is the viral production rate by infected cells;  $d_v$  is the viral clearance rate; and  $\lambda_y$  represents the rate at which latently infected cells activate to become productively infected cells.

The desired parameter  $r$  is the net growth rate of the infection at low viral loads, and in terms of the paper’s model, corresponds to:  $r = (\beta\lambda\gamma)/(d_x d_v) - d_y$ . To estimate the population-level posterior distribution for  $r$ , we used the paper’s reported posterior distributions of the basic parameters

for each patient. They reported these distributions as a list of 200,000 samples for each patient, from which we ignored the first 50,000 to allow convergence and used the remaining 150,000 as the individual-level posterior estimate for that patient. We then treated each posterior estimate of  $\log(r)$  as a random normal variable sampled from  $\mathcal{N}(\mu_i, \sigma_i)$ , where each patient's  $\mu_i$  is sampled from  $\mathcal{N}(\mu, \sigma)$ . Population parameters were estimated by maximum likelihood using the `mvmeta` library in R to be  $\mu = -0.398$  and  $\sigma = 0.194$ . The geometric mean of the corresponding lognormal distribution therefore corresponds to  $r = 0.4 \text{ d}^{-1}$ .

We confirmed these values with data from a separate treatment interruption trial<sup>9</sup>, using a least-squares fit to exponential growth of viral load shortly after rebound. This analysis again produced a geometric mean value of  $r = 0.4 \text{ d}^{-1}$ .

Another commonly used measure of viral fitness is the basic reproductive ratio  $R_0$ , which describes the average number of new cells infected by virus from a single actively infected cell. The growth rate and the basic reproductive ratio relate to each other as  $R_0 = r/d_y + 1$ , where  $d_y = b + d$ . Using the well-established average lifespan of infected CD4<sup>+</sup> T cells of 1 day (measured from viral load decay during ART<sup>63</sup>) a growth rate of  $r = 0.4$  corresponds to  $R_0 = 1.4$ . This basic reproductive ratio observed during ART interruption is consistently lower than that observed during acute infection (for which  $R_0 \approx 8 - 10$ )<sup>38:64</sup>, likely due to improved immune response.

## Estimation of $A$

We use the Luo *et al.*<sup>52</sup> data and a similar procedure to estimate  $A$ , the number of cells exiting the LR per day during fully suppressive ART, which is proportional to the initial residual viral load from which rebound begins. As this data included densely sampled viral load, but few measurements of CD4<sup>+</sup> T cells, the fitted parameters are most reliable in determining virological quantities, not cell counts. For example, the quantity  $(\gamma\lambda_y)/(d_v d_y)$ , which equals the residual viral load observed during fully suppressive therapy, should be particularly reliable. Additionally,  $d_y$  itself can be reliably determined the decay of viral load upon resumption of therapy<sup>63</sup>. Since it is the product of two reliable parameters, the rate of virus production during ART ( $\lambda_v = (\gamma\lambda_y)/d_v$ ) will also be a reliable quantity to estimate. In contrast, the cellular rate  $\lambda_y$  by itself may not be reliably estimated from this data. In fact, since Luo *et al.*<sup>52</sup> identified the sum  $x(t) + y(t)$  with *total* CD4 count, though the majority of CD4<sup>+</sup> T cells are not actually infectable, they should systematically report an *overestimate* of parameter  $\lambda_y$ . We therefore felt that  $A$  could be estimated more robustly by exploiting its proportionality to reliably estimated viral rates.

We estimate  $A$  in two steps: (1) use the posterior distributions reported by Luo *et al.*<sup>52</sup> to estimate the interpatient distribution for  $\lambda_v$ , which has units of copies of viral RNA per ml plasma, per day; (2) use separate observations regarding the ratio of cells to virus to scale  $\lambda_v$  to  $A$ , which has units of (total body) infected CD4<sup>+</sup> T cells per day. Analogously to the above estimation of  $r$ , we estimated a lognormal interpatient distribution for  $\lambda_v$ , obtaining log mean and standard deviation  $-1.469$  and  $0.991$ , respectively. To scale to  $A$ , we used the cell-to-virus ratio described above, which implies that  $A$  is  $(750 \text{ to } 3750) \times \lambda_v$ . To obtain a normal distribution for  $\log(A)$ , we treated the extremes of 750 and 3750 as estimates of the 95% CI for the ratio  $A/\lambda_v$ , adding  $\frac{1}{2}(\log(750) + \log(3750)) = 3.225$  to the mean and  $0.178$  to the standard deviation. The resulting distribution for  $A$  has log mean and standard deviation  $1.755$  and  $1.007$ , respectively.

## Estimation of $P_{Est}$

Measuring establishment probabilities in early stages of infections is a difficult task that requires frequent tracking of rare infection and extinction events in small populations (e.g., Hofacre *et al.*<sup>47</sup>). In contrast, even the most sensitive methods currently used to measure growth of HIV over time *in vitro* do not yet provide single-cell resolution<sup>45;65</sup>. In the absence of the relevant experimental results, we rely on population genetic models to estimate  $P_{Est}$ .

In the basic stochastic model described above, the establishment probability is mainly controlled by the parameter  $\lambda$ . For a fixed growth rate  $r$ , higher  $\lambda$  values correspond to lower values of  $P_{Est}$ . To generalize this concept to measures more common in population genetics, we reframe  $P_{Est}$  in terms of the quantity  $\rho$  – the ratio of the variance to the mean offspring number for a single cell. In our basic model,  $\rho = 1 + d\lambda/(b + d)$ , as the mean offspring number (basic reproductive ratio) is  $R_0 = b\lambda/(b + d)$  and the variance is  $(b(\lambda + (\lambda - R_0)^2) + dR_0^2)/(b + d)$ . Thus the establishment probability depends only on  $\rho$  and  $R_0$  and is the solution to

$$P_{Est}(R_0 + \rho - 1) = R_0(1 - e^{-P_{Est}(R_0 + \rho - 1)}). \quad (5)$$

Using this equation we can use estimates of  $\rho$  from various data sources to derive a range of reasonable values for  $P_{Est}$ .

The maximum establishment probability compatible with a fixed viral fitness  $R_0$  occurs when  $d = 0$ , which sets  $\rho = 1$  and results in  $P_{Est}$  being determined by the equation  $P_{Est} = 1 - e^{-P_{Est}R_0}$ . This equation corresponds to the establishment probability relationship derived by Pennings<sup>54</sup> and by Pearson *et al.*<sup>39</sup> for their “burst model.” Thus the  $\rho = 1$  limit sets a natural upper bound for  $P_{Est}$ .

To estimate a lower bound for  $P_{Est}$ , note that the ratio  $\rho$  measures the deviation of viral replication from a simple Poisson process. Population genetic studies have consistently measured the effective population size ( $N_e$ ) of an individual host’s infection as far smaller than the total number of cells infected with HIV ( $N$ )<sup>66</sup>, an indication of greater-than-Poisson variance in the replication process. The ratio  $N/N_e$  reports an order-of-magnitude estimate for  $\rho$ . Since equilibrium models of neutral diversity may not be applicable to a population such as HIV that undergoes frequent selective sweeps<sup>55;66</sup>, we consider two studies based on models of selective sweeps, both of which converge upon an estimate of  $N_e \approx 10^5$  for patients on either no treatment or ineffective treatment<sup>53;55</sup>. Using  $N \approx 10^8$  as a typical number of infected cells off therapy, these studies imply  $\rho \approx 10^3$ .

Some care is required to interpret this value in the context of our model: since these studies examined diverse, evolving viral populations, strains with low relative fitness tend to contribute negligibly to  $N_e$ <sup>55</sup>. In the small-infection regime of our model, however, relative fitness differences between strains are unimportant, and any virus that is viable (in the sense of absolute fitness,  $r > 0$  or  $R_0 > 1$ ) is capable of establishing a rebounding infection. The estimate  $\rho \approx 10^3$  is therefore a conservative upper bound for the true value of  $\rho$ , providing a lower bound for  $P_{Est}$ .

The most relevant estimate for the establishment probability to our knowledge comes from an analysis of the rate of emergence of drug-resistant variants from the latent reservoir during fully suppressive ART. This study calculated an “effective exit rate” from the latent reservoir of 5 productively infected cells per day<sup>54</sup>. When compared to our estimate of  $A$  centered at 57 cells per

day (1), this implies  $\rho \approx 57/5 = 11.4$ . We therefore use  $\rho \approx 10$  as the best parameter estimate and the center of sampled range.

To capture the current state of knowledge and represent our uncertainty, we use a lognormal distribution for  $\rho$  centered at 10 with 95% of the distribution falling between 1 and 100. After sampling  $r$  and  $\rho$ , we compute  $R_0$  (described above in the section on estimating  $r$ ) and then  $P_{Est}$  (using Eq. (5)). Interestingly, a recent experiment measured the *in vitro* establishment probability of a single cell productively infected with a reporter HIV strain to be 14%, which is at the 70th percentile of our distribution (G. Laird, personal communication).

## Analysis of alternate stochastic models

The basic five-parameter, two-variable model described above was used to derive the result that the outcomes of reservoir-reducing therapy depend only on four “key parameters,” which are not necessarily model-specific. To test the robustness of this claim to structural variation in model choice, we constructed a series of alternate stochastic models, some of which include more complex infection processes. For each model, we derived expressions relating the model-specific parameters to the four key parameters. We then simulated the model and compared it to simulations of the basic model with identical values of the key parameters. There is generally an infinite set of model-specific parameters satisfying the same key parameter values; where needed, we evaluated a few examples at the extremes of what could be considered biologically realistic. A full description of the models, parameters, and results is given in the Supplementary Materials. In summary, the alternate models were:

1. Constant burst model: Infected cells either die without infecting, or produce a fixed (integer) number of new infected cells.
2. Eclipse phase model: Upon reactivation from latency or new infection, cells enter an “eclipse phase” in which they cannot infect others (no virus is produced). Cells in the eclipse phase may either die or proceed to the productively infected phase, where they behave as infected cells do in the constant burst model. We consider a eclipse phase of average length 2 days with 1/6 chance of death.
3. Free virus model: We explicitly track the amount of free virus. Infected cells either die without infecting, or produce a fixed (integer) number of virions. Each virion may either be cleared or may infect a new cell. We consider viral burst sizes ranging from 10 to 1000 virions.
4. Homeostatic proliferation: Latently infected cells can either reactivate, die, or divide *without reactivating* giving rise to another latently infected cell. To test a reasonable biological limit of fast turnover, cells have an average lifespan of 6 months, but the net LR half-life is maintained at 44 months by proliferation.
5. Bursting homeostatic proliferation: Latently infected cells can either reactivate, die, or divide multiple times *without reactivating* giving rise to a burst of other latently infected cells. To

test a biological limit of fast turnover, cells have an average lifespan of 6 months, but the net LR half-life is maintained at 44 months by proliferation with 4 divisions leading to 16 cells.

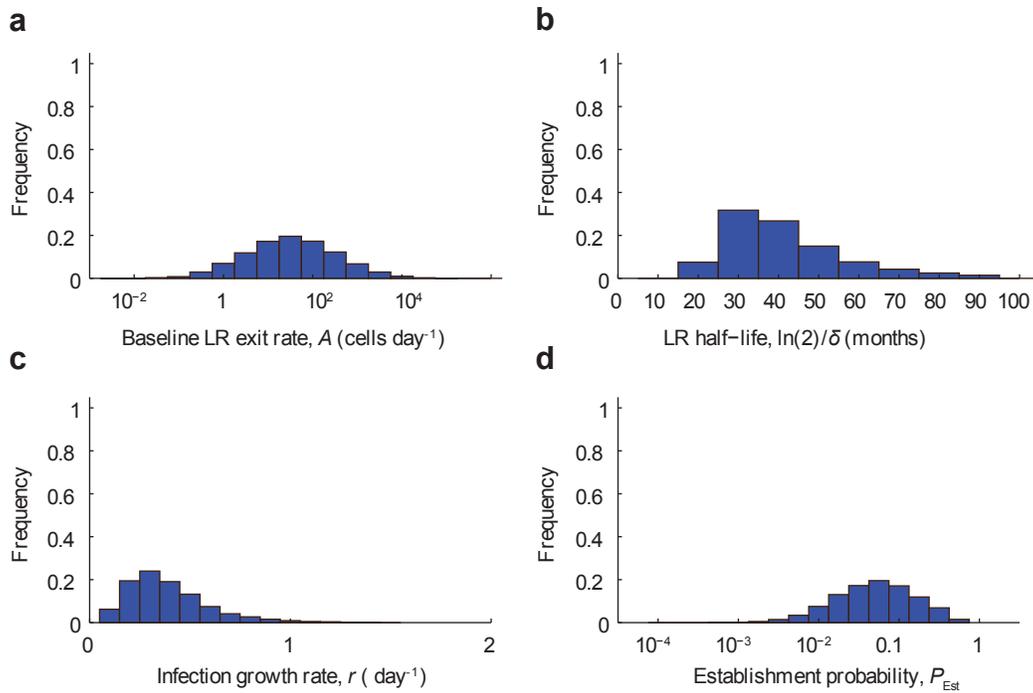
6. Expansion upon reactivation: When latently infected cells reactivate, they first proliferate multiple times *without infecting others*. Thus each reactivated latent cell results in multiple productively infected cells. We consider a maximum of 4 divisions leading to 16 newly reactivated cells.

## Simulation of the model

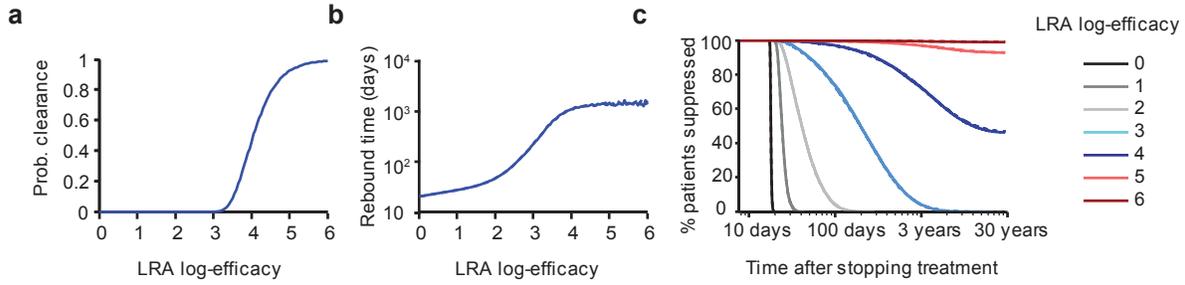
We use the Gillespie algorithm to track the number of latently and actively infected cells in a continuous time stochastic process. We start with an initial number of latent cells  $\mathcal{Z}(0) \sim \text{Binomial}(N_{LR}, q)$ , where  $N_{LR}$  is the pre-treatment reservoir size and  $q$  is the efficacy of LRA treatment (fraction of cells remaining). The initial number of actively infected cells  $\mathcal{Y}(0)$  is then chosen from a Poisson distribution with parameter  $a \mathcal{Z}(0)/d_y$  (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  $200 \text{ c ml}^{-1}$  (equivalent to  $\mathcal{Y} = 3 \times 10^5$  cells total) or until no active or latent cells remain. Because stochastic effects are important only for small  $\mathcal{Y}$ , we switch to faster deterministic numerical integration when  $\mathcal{Y}$  reaches a level where the probability of extinction is very low ( $< 10^{-4}$ ). 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 the model-specific parameters that are consistent with the sampled key parameters. In general, we use a pre-therapy LR size of  $N_{LR} = 10^6$  cells to get  $a = A/N_{LR}$ . We then have  $d_z = \delta - a$ . Using  $r$  and  $\rho$  along with  $d_y = d + b = 1 \text{ day}^{-1}$ , we can get  $\lambda, b, d$ , and  $P_{Est}$ . Consistent with our generating function analysis, we find that the specific values assumed for  $N_{LR}$  and  $d_y$  do not influence the results. For simulating other models, any other parameter assumptions are listed in the corresponding supplementary figure captions.

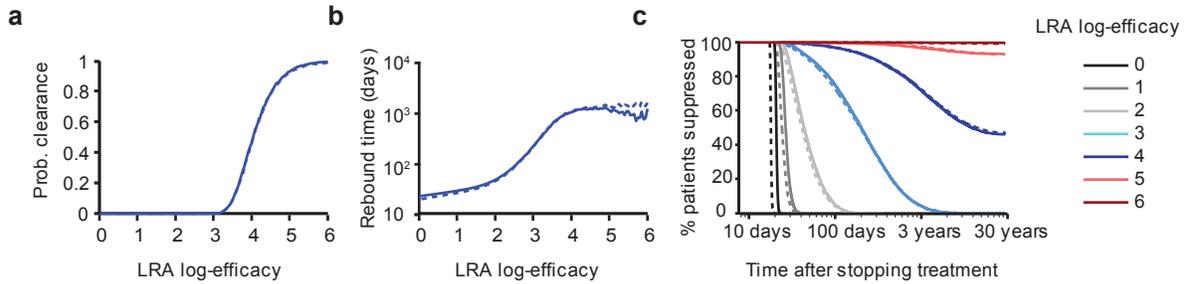
# Supplementary Figures



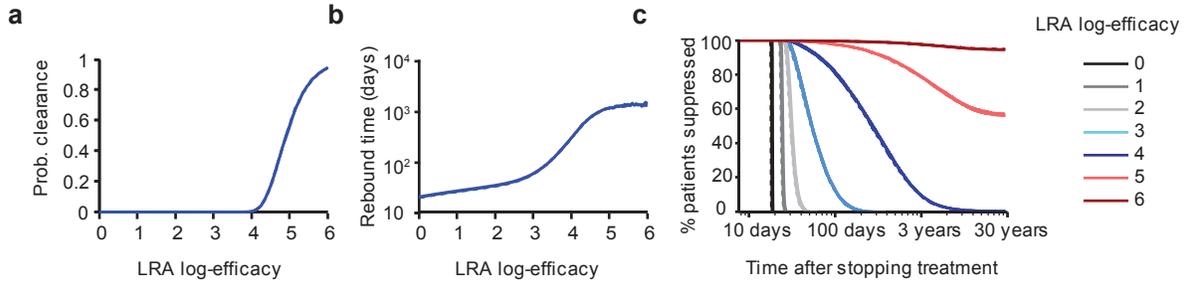
**Figure S1:** Parameter distributions described in Table 1 and used for results in Figs. 3 (uncertainty analysis) and 4 (target LRA efficacy)



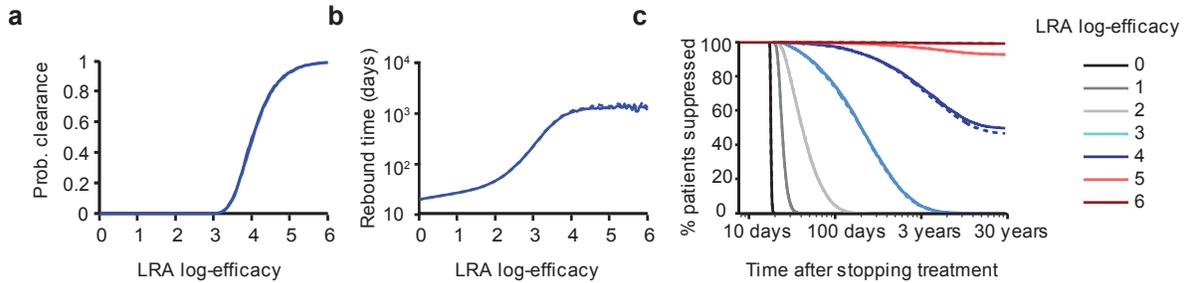
**Figure S2:** Comparing model predictions from the constant burst model (solid lines) with the basic (Poisson burst) model (dotted lines). For both models  $r = 0.4 \text{ d}^{-1}$ ,  $A = 57 \text{ cells}$ ,  $\delta = 5.23 \times 10^{-4} \text{ d}^{-1}$ ,  $P_{Est} = 0.07$ . For each model all patients have the same underlying viral dynamic parameters: constant burst  $b = 0.1273 \text{ d}^{-1}$ ,  $d = 0.8727 \text{ d}^{-1}$ ,  $\lambda = 11$ ; Poisson burst  $b = 0.137 \text{ d}^{-1}$ ,  $d = 0.863 \text{ d}^{-1}$ ,  $\lambda = 10.22$ ; both  $a = 5.7 \times 10^{-5} \text{ d}^{-1}$ ,  $d_z = 4.66 \times 10^{-4} \text{ d}^{-1}$ . a) 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. b) Median viral rebound times (logarithmic scale), among patients who do not clear the infection. c) 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  $10^4$  to  $10^5$  patients. Lack of visible dashed lines indicates both models give indistinguishable results.



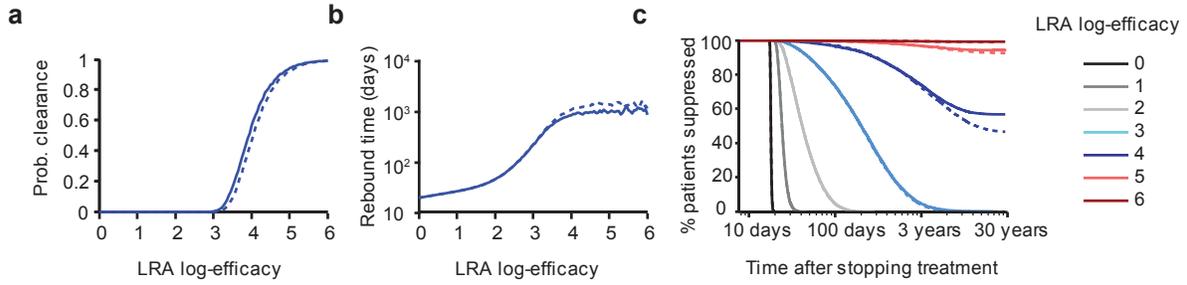
**Figure S3:** Comparing model predictions from the eclipse phase model (solid lines) with the basic (Poisson burst) model (dotted lines). For both models  $r = 0.4 \text{ d}^{-1}$ ,  $A = 57 \text{ cells}$ ,  $\delta = 5.23 \times 10^{-4}$ ,  $P_{Est} = 0.07$ . For each model all patients have the same underlying viral dynamic parameters: eclipse phase  $b = 0.0952$ ,  $d = 0.9048$ ,  $\lambda = 29$ ,  $e = 0.5 \text{ d}^{-1}$ ,  $f = 0.1 \text{ d}^{-1}$ ; Poisson burst  $b = 0.137 \text{ d}^{-1}$ ,  $d = 0.863 \text{ d}^{-1}$ ,  $\lambda = 10.22$ ; both  $a = 5.7 \times 10^{-5} \text{ d}^{-1}$ ,  $d_z = 4.66 \times 10^{-4} \text{ d}^{-1}$ . a) 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. b) Median viral rebound times (logarithmic scale), among patients who do not clear the infection. c) 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  $10^4$  to  $10^5$  patients. Lack of visible dashed lines indicates both models give indistinguishable results.



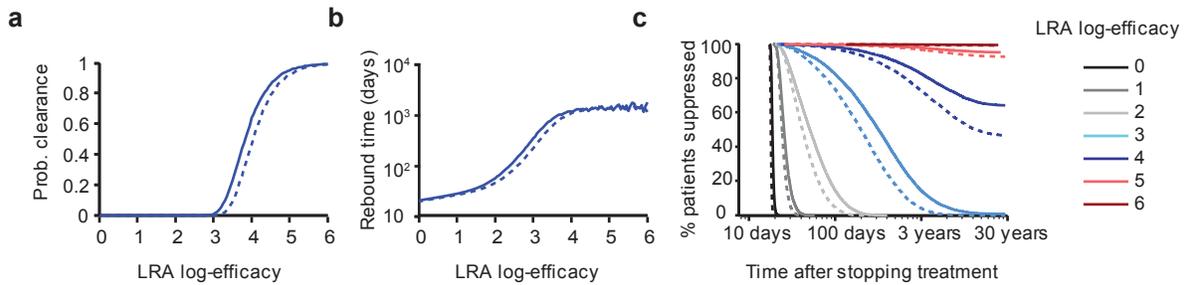
**Figure S4:** Comparing model predictions from two variations of the free virus model (solid lines) with the basic (Poisson burst) model (dotted lines). For both models  $r = 0.4 \text{ d}^{-1}$ ,  $A = 57 \text{ cells}$ ,  $\delta = 5.23 \times 10^{-4} \text{ d}^{-1}$ ,  $P_{Est} = 0.51$ . For each model all patients have the same underlying viral dynamic parameters: free virus 1)  $N = 10$ ,  $k = 0.92 \text{ d}^{-1}$ ,  $d = 0.08 \text{ d}^{-1}$ ,  $b = 3.6 \text{ d}^{-1}$ ,  $c = 20 \text{ d}^{-1}$ ; 2)  $N = 1000$ ,  $k = 0.997 \text{ d}^{-1}$ ,  $d = 0.0028 \text{ d}^{-1}$ ,  $b = 0.028 \text{ d}^{-1}$ ,  $c = 20$ ; Poisson burst  $b = 1$ ,  $d = 0$ ,  $\lambda = 1.4$ ; both  $a = 5.7 \times 10^{-5}$ ,  $d_z = 4.66 \times 10^{-4}$ . a) 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. b) Median viral rebound times (logarithmic scale), among patients who do not clear the infection. c) 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  $10^4$  to  $10^5$  patients. Lack of visible dashed lines and distinguishable solid lines indicates that the three models give indistinguishable results.



**Figure S5:** Comparing model predictions from the homeostatic proliferation model (solid lines) with the basic (Poisson burst) model (dotted lines). For both models  $r = 0.4 \text{ d}^{-1}$ ,  $A = 57 \text{ cells}$ ,  $\delta = 5.23 \times 10^{-4}$ ,  $P_{Est} = 0.069$ . For each model all patients have the same underlying viral dynamic parameters: homeostatic proliferation model  $p = 3.3 \times 10^{-3} \text{ d}^{-1}$ ,  $d_z = 3.8 \times 10^{-3} \text{ d}^{-1}$ ; Poisson burst  $d_z = 4.66 \times 10^{-4} \text{ d}^{-1}$ ; both  $b = 0.1346 \text{ d}^{-1}$ ,  $d = 0.8654 \text{ d}^{-1}$ ,  $\lambda = 10.4$ ,  $a = 5.7 \times 10^{-5} \text{ d}^{-1}$ . a) 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. b) Median viral rebound times (logarithmic scale), among patients who do not clear the infection. c) 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  $10^4$  to  $10^5$  patients. Lack of visible dashed lines indicates both models give indistinguishable results.



**Figure S6:** Comparing model predictions from the bursting homeostatic proliferation model (solid lines) with the basic (Poisson burst) model (dotted lines). For both models  $r = 0.4 \text{ d}^{-1}$ ,  $A = 57 \text{ cells}$ ,  $\delta = 5.23 \times 10^{-4}$ ,  $P_{Est} = 0.069$ . For each model all patients have the same underlying viral dynamic parameters: bursting homeostatic proliferation model  $p = 2.2 \times 10^{-4} \text{ d}^{-1}$ ,  $h = 4$ ,  $d_z = 3.8 \times 10^{-3} \text{ d}^{-1}$ ; Poisson burst  $d_z = 4.66 \times 10^{-4} \text{ d}^{-1}$ ; both  $b = 0.1346 \text{ d}^{-1}$ ,  $d = 0.8654 \text{ d}^{-1}$ ,  $\lambda = 10.4$ ,  $a = 5.7 \times 10^{-5} \text{ d}^{-1}$ . a) 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. b) Median viral rebound times (logarithmic scale), among patients who do not clear the infection. c) 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  $10^4$  to  $10^5$  patients. Lack of visible dashed lines indicates both models give indistinguishable results.



**Figure S7:** Comparing model predictions from the expansion upon reactivation model (solid lines) with the basic (Poisson burst) model (dotted lines). For both models  $r = 0.4 \text{ d}^{-1}$ ,  $A = 57 \text{ cells}$ ,  $\delta = 5.23 \times 10^{-4}$ ,  $P_{Est} = 0.069$ . For each model all patients have the same underlying viral dynamic parameters: expansion upon reactivation model  $m = 4$ ,  $a = 3.56 \times 10^{-6}$ ; Poisson burst  $a = 5.7 \times 10^{-5} \text{ d}^{-1}$ ; both  $b = 0.1346 \text{ d}^{-1}$ ,  $d = 0.8654 \text{ d}^{-1}$ ,  $\lambda = 10.4$ ,  $d_z = 4.66 \times 10^{-4} \text{ d}^{-1}$ . a) 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. b) Median viral rebound times (logarithmic scale), among patients who do not clear the infection. c) 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  $10^4$  to  $10^5$  patients.

# Supplementary Methods

## 1 Generating function analysis

### 1.1 Generating functions and clearance probability

Let  $f_0(\xi_0, \xi_1; t)$  and  $f_1(\xi_0, \xi_1; t)$  be probability generating functions for the process described in Methods starting from one latent and one active cell, respectively. Here,  $\xi_0$  and  $\xi_1$  are the dummy variables corresponding to the number of latent cells ( $\mathcal{Y}(t)$ ) and active cells ( $\mathcal{Z}(t)$ ), respectively, and  $t$  is time since treatment interruption. The backwards Kolmogorov equations<sup>57</sup> are given by the system of coupled ordinary differential equations

$$\begin{aligned}\frac{\partial f_0}{\partial t} &= a(f_1 - f_0) + d_z(1 - f_0), \\ \frac{\partial f_1}{\partial t} &= b(\exp[\lambda(f_1 - 1)] - f_1) + d(1 - f_1),\end{aligned}\tag{S1}$$

with boundary conditions  $f_0(\xi_0, \xi_1; 0) = \xi_0$  and  $f_1(\xi_0, \xi_1; 0) = \xi_1$ . The birth term  $\exp(\lambda(f_1 - 1))$  follows from the Poisson offspring distribution with parameter  $\lambda$ .

After LRA therapy, the initial reservoir size  $\mathcal{Z}(0)$  is Poisson-distributed with mean  $qN_{LR}$ . The initial residual viremia  $\mathcal{Y}(0)$  is determined by activation-death equilibrium during ART, and so it is Poisson-distributed with mean  $aqN_{LR}/(b + d)$ . The probability generating function corresponding to this initial condition is

$$f(\xi_0, \xi_1; t) = \exp \left[ qN_{LR} (f_0(\xi_0, \xi_1; t) - 1) + \frac{aqN_{LR}}{b + d} (f_1(\xi_0, \xi_1; t) - 1) \right].\tag{S2}$$

The fixed points of the differential equations (S1) give the extinction probabilities starting with one latent cell and one productive cell, respectively, denoted  $P_{Clr|\mathcal{Z}=1}$  and  $P_{Clr|\mathcal{Y}=1}$ . The extinction probability starting from the actual initial conditions, denoted  $P_{Clr}$ , then equals  $f(\xi_0, \xi_1; t)$  with the substitutions  $f_0(\xi_0, \xi_1; t) = P_{Clr|\mathcal{Z}=1}$  and  $f_1(\xi_0, \xi_1; t) = P_{Clr|\mathcal{Y}=1}$ :

$$P_{Clr} = \exp \left[ -aqN_{LR} (1 - P_{Clr|\mathcal{Y}=1}) \frac{b + d + a + d_z}{(b + d)(a + d_z)} \right] \approx \exp \left[ -\frac{aqN_{LR} (1 - P_{Clr|\mathcal{Y}=1})}{a + d_z} \right],\tag{S3}$$

where the approximation follows from the fact that rates controlling productive cells ( $b$  and  $d$ ) are much larger than rates controlling latent cells ( $a$  and  $d_z$ ).

### 1.2 Simplified branching process

Obtaining an explicit formula for rebound probability at a given time is not feasible in the above system. From here on we use a simplified branching process in which each birth event results in exactly two cells. The differential equation for  $f_1(\xi_0, \xi_1; t)$  is then

$$\frac{\partial f_1}{\partial t} = b (f_1(\xi_0, \xi_1; t)^2 - f_1(\xi_0, \xi_1; t)) + d(1 - f_1(\xi_0, \xi_1; t)), \quad (\text{S4})$$

while the equation for  $f_0(\xi_0, \xi_1; t)$  is unchanged. In this simplified process,  $P_{clr|y=1} = d/b$ . Parameters  $b$ ,  $d$ ,  $a$ , and  $d_z$  can be replaced with the parameters estimated in Table 1, using the relationships  $r = b - d$ ,  $P_{Est} = 1 - P_{clr|y=1} = 1 - d/b$ ,  $\delta = a + d_z$ , and  $A = aN_{LR}$ . Moreover, let  $A' = aqN_{LR}$ , the rate at which activated cells are produced, after therapy. The reservoir size  $N_{LR}$  then no longer appears in the expression for clearance probability:

$$P_{Clr} = \exp \left[ -\frac{A' P_{Est}}{\delta} \left( 1 + \frac{\delta P_{Est}}{r(2 - P_{Est})} \right) \right]. \quad (\text{S5})$$

Note that in the limit where active cell dynamics are much faster than latent cell dynamics ( $\delta/r \rightarrow 0$ ),  $P_{Clr}$  approaches  $\exp[-A' P_{Est}/\delta]$ , consistent with (S3).

### 1.3 Rebound time

The system with simplified replication dynamics (S4) can be solved explicitly:

$$\begin{aligned} f_0(\xi_0, \xi_1; t) &= 1 - a \frac{1 - e^{-\delta t}}{\delta} + a e^{-\delta t} \int_0^t e^{\delta \tau} f_1(\xi_0, \xi_1; \tau) d\tau, \\ f_1(\xi_0, \xi_1; t) &= \frac{(1 - P_{Est})(e^{rt}(1 - \xi_1) - 1) + \xi_1}{-(1 - P_{Est}) + e^{rt}(1 - \xi_1) + \xi_1}. \end{aligned} \quad (\text{S6})$$

Let  $f(\xi_0, \xi_1; t)$  be the generating function for the initial conditions described in 1.1, above. As rebound is defined only by the number of actively infected cells, it suffices to analyze the marginal generating function  $f(1, \xi_1; t)$ . From here on, the term  $\xi_0$  is suppressed and  $\xi_1$  is replaced with  $\xi$ :

$$f(\xi; t) = \exp \left[ qN_{LR}(f_0(\xi; t) - 1) + \frac{aqN_{LR}P_{Est}}{r(2 - P_{Est})}(f_1(\xi; t) - 1) \right]. \quad (\text{S7})$$

Assuming that active cell dynamics are much faster than latent cell dynamics (small  $\delta/r$ , to first-order), an explicit formula can be given for the expansion of the generating function  $f(\xi; t)$  at  $\xi = 0$ . This expansion yields  $P(\mathcal{Y}, t)$ , the probability that there are exactly  $\mathcal{Y}$  actively infected cells at time  $t$  (see 1.4.1, below, for derivation):

$$\begin{aligned} P(\mathcal{Y}, t) &\approx \exp \left[ -A' P_{Est} \left( \omega_1 + \frac{1 - e^{\delta t} \left( 1 + \delta \int_0^t g(0; \tau) d\tau \right)}{\delta P_{Est}} \right) \right] \\ &\times \frac{1}{\mathcal{Y}!} \left( \frac{e^{rt} - 1}{e^{rt} - 1 + P_{Est}} \right)^{\mathcal{Y}} \\ &\times (\omega_2)_{\mathcal{Y}} M \left( -\mathcal{Y}, \omega_2, -\frac{A' P_{Est}^2}{2 - P_{Est}} \omega_1 \right), \end{aligned} \quad (\text{S8})$$

where  $\omega_1 = (e^{rt}P_{Est}) / (r(e^{rt} - 1)(e^{rt} - 1 + P_{Est}))$ ,  $\omega_2 = (A'P_{Est}e^{-\delta t}) / r$ ,  $g(t) = e^{\delta t}(e^{rt} - 1)(1 - P_{Est}) / (e^{rt} - 1 + P_{Est})$ ,  $(\cdot)_y$  is the Pochhammer symbol (rising factorial), and  $M$  is Kummer's confluent hypergeometric function. Note that unlike with standard techniques for inverting generating functions<sup>37;67</sup>, the integrand on the first line of (S8) is not a highly oscillatory function, and so numerical computation is efficient.

Let  $P(\geq N_{Reb}, t)$  be the probability of rebound, defined as there being at least  $N_{Reb}$  actively infected cells at time  $t$ . Direct computation of this quantity as  $1 - \sum_{\mathcal{Y}=0}^{N_{Reb}-1} P(\mathcal{Y}, t)$  is relatively slow. In 1.4.2, below, we show that for realistically large  $N_{Reb}$ , this probability can be efficiently computed as

$$P(\geq N_{Reb}, t) \approx (1 - P_{Clr}) \frac{\int_0^t P(N_{Reb}, \tau) d\tau}{\int_0^\infty P(N_{Reb}, \tau) d\tau}. \quad (\text{S9})$$

A script is provided at <http://www.danielrosenbloom.com/reboundtimes> using this formula for rapid computation of survival curves.

## 1.4 Additional derivations

### 1.4.1 Expansion of the probability generating function

The generating function (S7) can be written

$$f(\xi; t) = \exp \left[ k_1 + b_1 g_1(\xi; t) + b_2 \int_0^t g_2(\xi; \tau) d\tau \right], \quad (\text{S10})$$

where

$$\begin{aligned} k_1 &= - \left( \frac{A'}{\delta} \right) (1 - e^{-\delta t}), \\ b_1 &= - \frac{A'P_{Est}^2 e^{rt}}{r(2 - P_{Est})}, \\ b_2 &= A' e^{-\delta t}, \\ g_1(\xi; t) &= \frac{1 - \xi}{(e^{rt} - 1)(1 - \xi) + P_{Est}}, \\ g_2(\xi; t) &= e^{\delta t} \frac{(e^{rt} - 1)(1 - \xi)(1 - P_{Est}) + P_{Est}\xi}{(e^{rt} - 1)(1 - \xi) + P_{Est}}. \end{aligned} \quad (\text{S11})$$

The derivatives of  $g_1(\xi; t)$  and  $g_2(\xi; t)$  at  $\xi = 0$  can be written explicitly:

$$\begin{aligned}
g_1(0; t) &= \frac{1}{e^{rt} - 1 + P_{Est}}, \\
g_2(0; t) &= e^{\delta t} \frac{(e^{rt} - 1)(1 - P_{Est})}{e^{rt} - 1 + P_{Est}}, \\
\left. \frac{\partial^j g_1(\xi; t)}{\partial \xi^j} \right|_{\xi=0} &= -j! \frac{(e^{rt} - 1)^{j-1} P_{Est}}{(e^{rt} - 1 + P_{Est})^{j+1}}, \\
\left. \frac{\partial^j g_2(\xi; t)}{\partial \xi^j} \right|_{\xi=0} &= j! e^{(r+\delta)t} \left( \frac{e^{rt} - 1}{e^{rt} - 1 + P_{Est}} \right)^{j-1} \left( \frac{P_{Est}}{e^{rt} - 1 + P_{Est}} \right)^2,
\end{aligned} \tag{S12}$$

for all  $j > 0$ .

Note that  $g_2(\xi; t)$  and its derivatives contribute to the generating function (S10) only via integration over  $\tau$  from 0 to  $t$ . In the limit of large  $\tau$ , the derivative  $\left. \frac{\partial^j g_2(\xi, \tau)}{\partial \xi^j} \right|_{\xi=0}$  approaches  $j! e^{(-r+\delta)\tau} P_{Est}^2$  for any  $j > 0$ . Assuming that active cell dynamics are faster than latent cell dynamics,  $\delta$  is much smaller than  $r$ , and so this derivative vanishes as  $r\tau$  grows. It therefore contributes meaningfully to the integral only at small  $r\tau$ , a regime in which  $e^{\delta t} \approx 1$ . For the purposes of calculating the generating function expansion, we therefore may safely ignore  $\delta$  in the derivatives of  $g_2(\xi; t)$ , arriving at the approximation

$$\left. \frac{\partial^j g_2(\xi, t)}{\partial \xi^j} \right|_{\xi=0} \approx j! e^{rt} \left( \frac{e^{rt} - 1}{e^{rt} - 1 + P_{Est}} \right)^{j-1} \left( \frac{P_{Est}}{e^{rt} - 1 + P_{Est}} \right)^2 \quad (\delta \ll r \text{ and } j > 0). \tag{S13}$$

Let  $P(\mathcal{Y}, t)$  denote the probability that  $\mathcal{Y}$  active cells are present at time  $t$ . This probability equals the coefficient of  $\xi^{\mathcal{Y}}$  in the expansion of  $f(\xi; t)$  around  $\xi = 0$ , which is  $\frac{1}{\mathcal{Y}!} \left. \frac{\partial^{\mathcal{Y}} f(\xi; t)}{\partial \xi^{\mathcal{Y}}} \right|_{\xi=0}$ . Using approximation (S13), the expansion is

$$\begin{aligned}
P(\mathcal{Y}, t) &\approx \frac{1}{\mathcal{Y}! (e^{rt} - 1 + P_{Est})^{2\mathcal{Y}}} e^{k_1 + b_1/(e^{rt} - 1 + P_{Est}) + b_2 \int_0^t g_2(0; \tau) d\tau} \\
&\times \sum_{i=0}^{\mathcal{Y}} \binom{\mathcal{Y}}{i} (-b_1 P_{Est})^i \left( (e^{rt} - 1)(e^{rt} - 1 + P_{Est}) \right)^{\mathcal{Y}-i} \\
&\quad \left( \frac{b_2 P_{Est}}{r} + i \right)_{\mathcal{Y}-i}.
\end{aligned} \tag{S14}$$

Expression (S8) in the previous section is a more convenient formulation of this probability in terms of a hypergeometric polynomial.

#### 1.4.2 Efficient computation of rebound probability

Rebound probability  $P(\geq N_{Reb}, t)$  is the probability that there are at least  $N_{Reb}$  active cells at time  $t$ . Define  $P(\geq N_{Reb}, \infty)$  as the probability that rebound *ever* occurs. Since realistic values of

$N_{Reb}$  are large (many thousands of infected cells),  $P(\geq N_{Reb}, \infty)$  is approximately one minus the extinction probability, given in (S5).

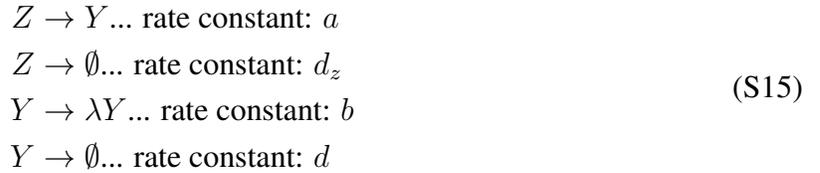
Also since  $N_{Reb}$  is large, we can assume that, once this rebound level is achieved, the number of active cells is almost always increasing; that is, decreases are rare and short-lived. The probability  $P(\geq N_{Reb}, t)$  is therefore nearly equal to the probability that there were *ever* exactly  $N_{Reb}$  cells sometime prior to  $t$ . Define  $E(N_{Reb}, t) = \int_0^t P(N_{Reb}, \tau) d\tau$ , the expected duration for which exactly  $N_{Reb}$  cells existed, prior to  $t$ . This expectation can be decomposed  $E(N_{Reb}, t) \approx P(> N_{Reb}, t)D(N_{Reb})$ , where  $D(N_{Reb})$  is the expected duration that the number of active cells stays at  $N_{Reb}$  conditional upon it reaching  $N_{Reb}$ . Again by large  $N_{Reb}$ , this duration is insensitive to latent cell dynamics and so is independent of time. Similarly, let  $E(N_{Reb}, \infty) = \int_0^\infty P(N_{Reb}, t) d\tau$ , which is approximately  $P(> N_{Reb}, \infty)D(N_{Reb})$ . The desired probability is therefore  $P(\geq N_{Reb}, t) \approx P(> N_{Reb}, \infty)E(N_{Reb}, t)/E(N_{Reb}, \infty) \approx (1 - P_{Clr}) E(N_{Reb}, t)/E(N_{Reb}, \infty)$ , as given in equation (S9).

## 2 Alternate stochastic models

### 2.1 Constant burst model

**Summary:** Infected cells either die without infecting, or produce a fixed (integer) number of new infected cells.

**Process:**



$Y$  and  $Z$  are individual actively or latently infected cells. An actively infected cell can either die (at rate  $d$ ) or produce a burst of virions (at rate  $b$ ) that results in the infection of  $\lambda$  other cells, where  $\lambda$  is an integer. After a burst event, the original cell dies. Free virus is not explicitly tracked, but assumed to be at a level proportional to the number of actively infected cells. Latently infected cells can either die (at rate  $d_z$ ) or reactivate and become productively infected (at rate  $a$ ). All rates have units of  $d^{-1}$ , and  $\lambda$  is unitless.

**Parameters:**

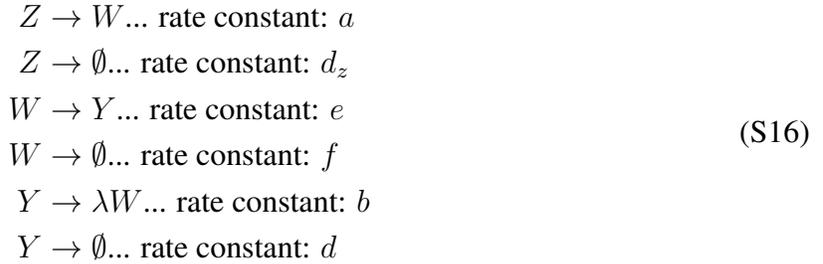
- $r = b\lambda - (b + d)$
- $A = aN_{LR}$
- $\delta = d_z + a$

- $P_{Est}$  is the solution to  $R_0(1 - (1 - P_{Est})^\lambda) - \lambda P_{Est} = 0$
- $R_0 = \frac{b\lambda}{d_y}$ ,  $d_y = (b + d)$

## 2.2 Eclipse phase model

**Summary:** Upon reactivation from latency or new infection, cells enter an “eclipse phase” in which they cannot infect others (no virus is produced). Cells in the eclipse phase may either die or proceed to the productively infected phase, where they behave according to infected cells in the constant burst model. We consider an eclipse phase of 2-day average length with 1/6 chance of death.

**Process:**



$Y$  and  $Z$  are individual productively or latently infected cells.  $W$  is an early “eclipse” phase cell. A productively infected cell can either die (at rate  $d$ ) or produce a burst of virions (at rate  $b$ ) that results in the infection of  $\lambda$  eclipse phase cells, where  $\lambda$  is an integer. After a burst event, the original cell dies. Free virus is not explicitly tracked, but assumed to be at a level proportional to the number of productively infected cells. Eclipse phase cells can either die (at rate  $f$ ) or proceed to the productively infected phase (at rate  $e$ ). Latently infected cells can either die (at rate  $d_z$ ) or reactivate and become eclipse phase cells (at rate  $a$ ). All rates have units of  $d^{-1}$ , and  $\lambda$  is unitless.

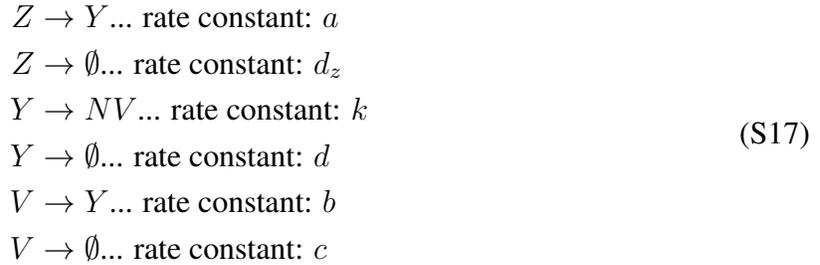
**Parameters**

- $r = (1/2)(-(d_w + d_y) + \sqrt{(d_w + d_y)^2 + 4d_w d_y (R_0 - 1)})$
- $A = aN_{LR}$
- $\delta = d_z + a$
- $P_{Est}$  is the solution to  $R_0(1 - (1 - P_{Est})^\lambda) - \lambda P_{Est} = 0$  (this is the establishment probability starting from a single cell in the eclipse phase. For a single productively infected cell,  $P_{Est}^y = \frac{d_w}{e} P_{Est}$ ).
- $R_0 = \frac{be\lambda}{d_w d_y}$ ,  $d_w = e + f$ ,  $d_y = b + d$

## 2.3 Free virus model

**Summary:** We explicitly track the amount of free virus. Infected cells either die without infection, or produce a fixed (integer) number of virions. Each virion may either be cleared or infect a new cell.

**Process:**



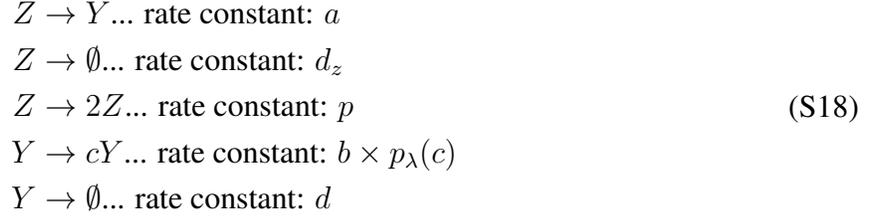
$Y$  and  $Z$  are individual actively or latently infected cells.  $V$  is a free virion. An actively infected cell can either die (at rate  $d$ ) or produce a burst of virions (at rate  $k$ ) that results in the production of  $N$  virions, where  $N$  is an integer. After a burst event, the original cell dies. Each free virion can either be cleared (at rate  $c$ ) or infect another cell (at rate  $b$ ). Latently infected cells can either die (at rate  $d_z$ ) or reactivate and become productively infected (at rate  $a$ ). All rates have units of  $\text{day}^{-1}$ , and  $N$  is unitless.

**Parameters:**

- $r = (1/2)(-(d_v + d_y) + \sqrt{(d_v + d_y)^2 + 4d_v d_y (R_0 - 1)})$
- $A = aN_{LR}$
- $\delta = d_z + a$
- $P_{Est} = \frac{d_v}{b} P_{Est}^v$ , where  $P_{Est}^v$  is the solution to  $R_0(1 - (1 - P_{Est}^v)^N) - NP_{Est}^v = 0$  ( $P_{Est}$  this is the establishment probability starting from a single cell and  $P_{Est}^v$  is the establishment probability starting from a single virion).
- $R_0 = \frac{kbN}{d_v d_y}$ ,  $d_v = b + c$ ,  $d_y = k + d$

## 2.4 Homeostatic proliferation

**Summary:** Latently infected cells can either reactivate, die, or divide *without reactivating* giving rise to another latently infected cells.

**Process:**

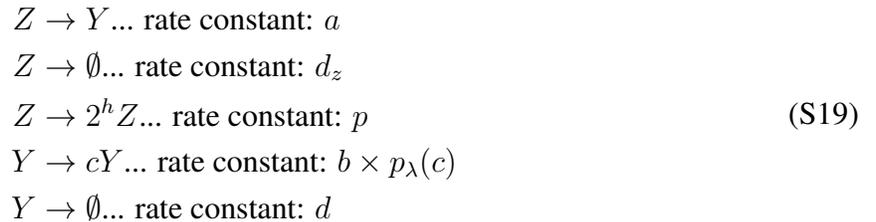
$Y$  and  $Z$  are individual actively or latently infected cells. An actively infected cell can either die (at rate  $d$ ) or produce a burst of virions (at rate  $b$ ) that results in the infection of  $c$  other cells, where  $c$  is a Poisson-distributed random variable with parameter  $\lambda$ ,  $p_\lambda(c) = \frac{\exp(-\lambda)\lambda^c}{c!}$ . After a burst event, the original cell dies. Free virus is not explicitly tracked, but assumed to be at a level proportional to the number of actively infected cells. Latently infected cells can either die (at rate  $d_z$ ), reactivate and become productively infected (at rate  $a$ ), or divide *without reactivating* giving rise to another latently infected cell. All rates have units of  $\text{day}^{-1}$ , and  $\lambda$  is unitless.

**Parameters:**

- $r = b\lambda - (b + d)$
- $A = aN_{LR}$
- $\delta = d_z + a - p$
- $P_{Est}$  is the solution to  $R_0(1 - e^{-\lambda P_{Est}}) - \lambda P_{Est} = 0$
- $R_0 = \frac{b\lambda}{d_y}$ ,  $d_y = (b + d)$

**2.5 Bursting homeostatic proliferation**

**Summary** Latently infected cells can either reactivate, die, or divide multiple times *without reactivating* giving rise to a burst of other latently infected cells.

**Process:**

$Y$  and  $Z$  are individual actively or latently infected cells. An actively infected cell can either die (at rate  $d$ ) or produce a burst of virions (at rate  $b$ ) that results in the infection of  $c$  other cells, where  $c$  is a Poisson-distributed random variable with parameter  $\lambda$ ,  $p_\lambda(c) = \frac{\exp(-\lambda)\lambda^c}{c!}$ . After a

burst event, the original cell dies. Free virus is not explicitly tracked, but assumed to be at a level proportional to the number of actively infected cells. Latently infected cells can either die (at rate  $d_z$ ), reactivate and become productively infected (at rate  $a$ ), or divide  $h$  times *without reactivating* giving rise to  $2^h$  latently infected cells (where  $h$  is an integer). All rates have units of  $\text{day}^{-1}$ , and  $\lambda$  and  $h$  are unitless.

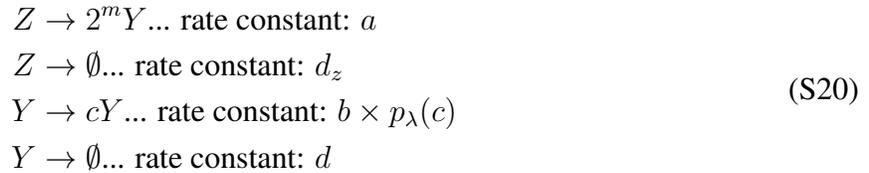
**Parameters:**

- $r = b\lambda - (b + d)$
- $A = aN_{LR}$
- $\delta = d_z + a - p(2^h - 1)$
- $P_{Est}$  is the solution to  $R_0(1 - e^{-\lambda P_{Est}}) - \lambda P_{Est} = 0$
- $R_0 = \frac{b\lambda}{d_y}$ ,  $d_y = (b + d)$

## 2.6 Expansion upon reactivation

**Summary:** When latently infected cells reactivate, they first proliferate multiple times *without infecting others*.

**Process:**



$Y$  and  $Z$  are individual actively or latently infected cells. An actively infected cell can either die (at rate  $d$ ) or produce a burst of virions (at rate  $b$ ) that results in the infection of  $c$  other cells, where  $c$  is a Poisson-distributed random variable with parameter  $\lambda$ ,  $p_\lambda(c) = \frac{\exp(-\lambda)\lambda^c}{c!}$ . After a burst event, the original cell dies. Free virus is not explicitly tracked, but assumed to be at a level proportional to the number of actively infected cells. Latently infected cells can either die (at rate  $d_z$ ), or reactivate (at rate  $a$ ) and proliferate  $m$  times to produce  $2^m$  actively infected cells (where  $m$  is an integer). All rates have units of  $\text{day}^{-1}$ , and  $\lambda$  and  $m$  are unitless.

**Parameters:**

- $r = b\lambda - (b + d)$
- $A = a2^m N_{LR}$
- $\delta = d_z + a$

- $P_{Est}$  is the solution to  $R_0(1 - e^{-\lambda P_{Est}}) - \lambda P_{Est} = 0$
- $R_0 = \frac{b\lambda}{d_y}$ ,  $d_y = (b + d)$