Katherine Cohen Senior Thesis

Acknowledgements

I would like to thank my thesis advisor, Alison Hill, for being an incredible mentor and friend in the process of writing this thesis. I could not have accomplished this without her guidance, patience, and commitment to this project. I would also like to thank my parents, roommates, and friends for their love, support, and words of encouragement.

Table of Contents

Introduction

The global pandemic now known as Acquired Immunodeficiency Syndrome (AIDS) was first recognized as a new disease in 1981. Since then, more than 70 million people have contracted the disease and about 35 million people have died of what was ultimately recognized as its causative agent: Human Immunodeficiency Virus (HIV) (UNAIDS, 2017). The HIV pandemic has proven one of the most devastating infectious disease outbreaks to arise in recent history. It likely would have continued to have a sweeping death toll had novel antiretroviral treatments not been developed and found effective. Though the world is still without a true cure, successful treatment of HIV began with with the development of the first working reverse transcriptase inhibitor, zidovudine (Broder, 2010). To understand zidovudine's preliminary success, it is first necessary to understand the virology of HIV (reviewed in Nowak & May, 2000).

HIV targets CD4+ T cells, the helper T cells that are crucial to the human adaptive immune response. HIV enters CD4+ T cells through fusion of its own viral membrane with the T cell membrane. It does so by utilizing surface proteins to bind to the CD4 receptor, as well as one of two coreceptors: CCR5 or CXCR4. This happens via interactions between CD4, coreceptor, and an HIV envelope protein known as gp-120. Once inside the cell, reverse transcriptase - an enzyme encoded by the virus's genome and carried in the viral particle - is responsible for the part of the life cycle key to all retroviruses: converting the viral RNA into a DNA form. The viral DNA is then inserted into the host genome using the virus's integrase protein, and host cell machinery is then used to transcribe and translate the virus genome to make new copies of the

genome and viral proteins. The infected T cell is transformed into a factory for new HIV virions, with potentially thousands of new HIV virus particles budding off from its surface. In the process of budding off from the infected CD4+ T cell, the cell membrane becomes the HIV viral envelope, encapsulating the newly transcribed viral genome and containing on its surface all membrane proteins necessary to continue the infection of future T cells.

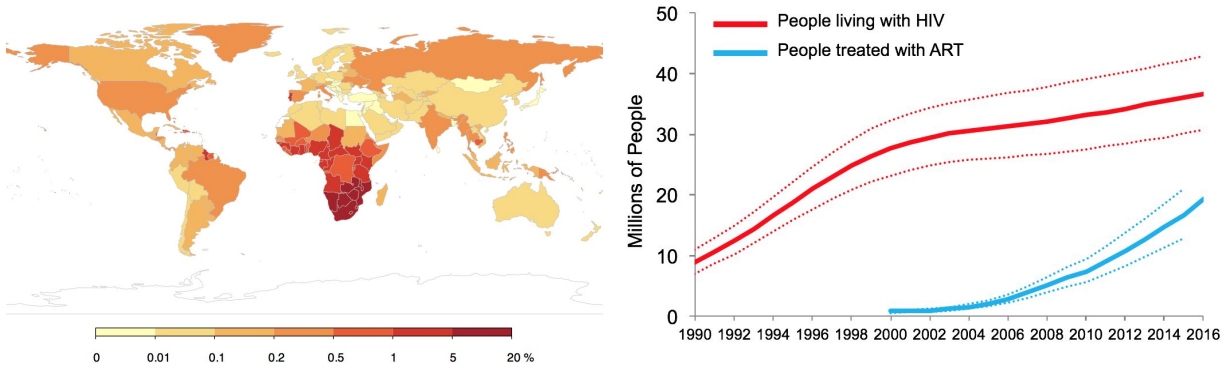


Figure 1: Global trends in HIV/AIDS. a) Map of the 2016 prevalence of HIV infection around the world. b) Longitudinal trends in the prevalence of HIV and the availability of antiretroviral therapy (Data from www.aidsinfo.unaids.org) .

In the process of reproducing, HIV takes a huge toll on the host CD4+ T cells, and thus on the immune system of the host organism. It does so by causing T cells to die either as a direct result of viral production or due to subsequent targeting by the organism's own cytolytic anti-viral immune response. Both pathways of T cell death leaves the host with a shortage of immune cells and thus in a severely immunodeficient state, leaving them incapable of fighting off any secondary infection. It is this immunodeficient state that arises from prolonged HIV infection and subsequent destruction of CD4+ T cells we call AIDS, and is how HIV ultimately kills (reviewed in Cummins et al 2014).

Returning to the discussion of HIV treatment and steps towards finding a cure, the drug previously mentioned, zidovudine, is a reverse transcriptase inhibitor, a common class of drug

for treatment of retroviruses such as HIV. The primary mechanism of action of a reverse transcriptase inhibitor is preventing the retrovirus from carrying out one functions crucial to its life cycle: utilizing reverse transcriptase to convert its own viral RNA into DNA to then be inserted into the host cell genome. Though this treatment proved initially to be quite effective in the short term window of just a few months, it faltered in that the high mutation rate of HIV enabled rapid development of resistance to the drug (Larder et al 1989, Nowak & McLean 1992). By 1995, a drug delivery strategy that combined two reverse transcriptase inhibitors and a protease inhibitor for the first time made it so that HIV diagnosis was not a guaranteed death sentence, extending the lives of those infected for many years (Ho et al, 1995, Jansson et al 2012).

Despite improvements in recent years of the state of HIV treatment, the world is not yet rid of HIV. In 2016, in fact, it was estimated that there were 36.7 million people worldwide living with HIV infection. Only about half of these individuals are currently on antiretroviral drugs, due to the significant cost and logistical hurdles to providing therapy to individuals in resource limited settings (UNAIDS 2017). An even more pressing challenge is that even those patients that are receiving ART treatment are required to continuously receive treatment forever. If they stop taking the drug the infection will relapse, meaning that these patients, despite receiving treatment, are still not cured of the virus (Eisele & Siliciano, 2012). HIV cannot be cured the same way that most viruses are cured because of its ability to lay dormant within immune cells for an extended amount of time. These "latent" HIV are not targeted by ART drugs because they are not actively in the process of replicating their viral genome. Thus upon the cessation of treatment, latent HIV remains unharmed by the drug, and is free to awaken at any

time to restart the infection. Almost universally, infection rebounds to pre-ART levels within a few weeks of stopping ART. Because of this unique property of the virus, we are still without a true cure for HIV. This means that the scientific community is still tasked with thinking about new ways to understand and ultimately combat the virus beyond our current understanding (Deeks et al, 2016).

Many different approaches are being considered to provide a permanent cure for HIV (Eisele & Siliciano, 2012). One idea, often referred to as a sterilizing cure, is to rid the body of all remaining latent virus so that nothing is left to restart the infection if ART is stopped. Another approach, often called a "functional cure" is to instead equip the body with the ability to control any remaining virus, without lifelong ART. The types of therapies being investigated to achieve one or both of these goals include latency-reversing drugs, immunotherapies (such as innate immune stimulators, checkpoint inhibitors, therapeutic vaccines, monoclonal antibodies, chimeric antigen receptor T cells), or gene therapy (e.g. HIV-resistant T cells, excision of provirus). In general, the idea is that these new therapies would be given in addition to traditional ART, with the hope that when all therapy is eventually stopped, viral rebound does not occur. However, in all likelihood, initially tested therapies will be imperfect, and rebound may still occur, but perhaps with altered kinetics - a longer delay, a slower growth rate, or reaching an eventual lower viral level. To date, it is impossible to predict how and when an individual will rebound, since it is very difficult to actually measure the levels of latent virus, and measuring the strength of the immune system against HIV is also complicated by an incomplete understanding of the mechanism of potential immune control. It is also currently unknown how the severity of

acute infection relates to the severity of rebound, since it is very difficult to diagnose humans during acute infection.

The goal of this project was to characterize the kinetics of HIV infection during acute infection (before ART) as well as during rebound (when ART is stopped), in the presence and absence of an new therapy being investigated for its potential to permanently cure the infection. To do this, we will use mathematical models which describe the interaction between the virus and the host during infection, and use data from a pre-clinical trial to inform the model parameters. Mathematical models have a long history of informing the study of infectious diseases, particularly at the level of describing the spread of epidemics in a population (reviewed in Anderson & May, 1990), but also in describing the spread of viral infections between cells in the body (reviewed in Nowak & May, 2000). For HIV, these so-called "viral dynamics" models have helped researchers understand phenomena such as the characteristic shape of viral load curves (from initial exponential increase, to peak, to lower set point equilibrium), the role of the target cell limitation vs the immune system in controlling infection, the effect of antiretroviral therapy on viral loads, the evolution of drug resistance, and the dynamics of creating and reactivation of latently infected cells (reviewed in Perelson & Ribeiro, 2013; Hill, 2018). In this thesis, I present a simple set of ordinary differential equations that can describe HIV infection and develop a method for fitting this model to experimental viral load data. The objective was to do obtain estimates for and analyze these biologically meaningful parameters during acute infection and rebound, with and without the addition of a new immunotherapy during ART. I evaluated the fitting method that I developed by testing the ability to to recover parameters in simulated data under increasingly non-ideal circumstances, including infrequent sampling,

measurement error, unobserved variables, and censored values. Once confirmed, I implemented the method on real viral load data to estimate the parameters of the model in each individual. I finally used the results of these fits to understand the effect of treatment, as well as to attempt to understand the differences between initial viral infection and infection post cessation of treatment.

Methods

Basics of viral dynamics model

Mathematical modeling is a tool can be used to understand the expected dynamics of HIV given particular assumptions about the biological mechanisms of viral infection. "Viral dynamics" models can track the kinetics of interactions between free viral particles, host target cells, anti-viral immune responses, and even drug treatment drug in an individual host.

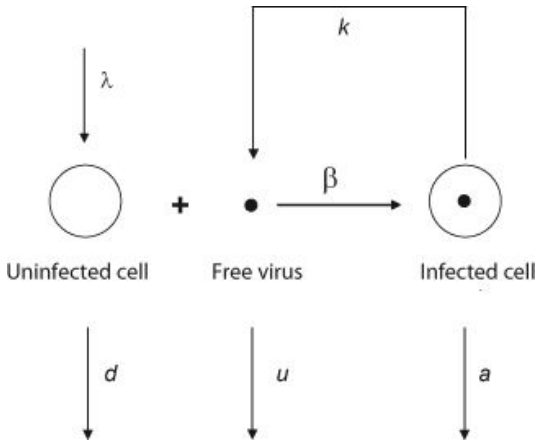


Figure 2: Schematic of the basic viral dynamics model. Parameters, variables, and reactions occuring in the model are described in the text.

The basic version of the viral dynamics model, which I will use throughout this thesis, describes the dynamics of uninfected and infected target cells and free virions, and has proven effective in modeling HIV and many other viral infections (e.g. Dahari et al 2009, Chatterjee et al 2013, Murillo et al 2013, Clapham et al 2014). The variables tracked in the model are x, the concentration of uninfected target cells, y, concentration of infected cells, and v, the concentration of free viral particles (Table 1). Since HIV infects T cells it is mainly an infection of the blood and lymph, and we assume the infection is homogeneous and well mixed in these tissues.

Table 1: Parameters and variables of the basic viral dynamics model

Parameter/Variable Name	Description	Units
x	Concentration of uninfected cells	cell/mL
y	Concentration of infected cells	cell/mL
v	Concentration of viral particles	copies/mL
λ	Rate of replenishment for uninfected cells	(cells/mL) day^{-1}
β	Rate of infection	$(copies/mL)^{-1}$ day^{-1}
k	Rate at which infected cells produce free virions	(virions/cell) day^{-1}
d	Death rate of uninfected cells	day^{-1}
a	Death rate of infected cells	day^{-1}
u	Clearance rate of free virus	day^{-1}

Similar to how chemical reactions between agents proceed, we can think about the dynamics between cells, both uninfected and infected, and virus as processes occurring at rates proportional to their abundance. The most important part of this model is the term governing the

rate of new infections, which depends on the abundance of both uninfected target cells (x) and abundance of virus (v), and on the overall rate at which productive infection occurs. Thus this rate of infection term is given by $\beta x v$, where the infectivity parameter β accounts for a number of rate-determining properties in this interaction, including rate of contact between virus and uninfected cell, rate of viral entry, as well as rate and probability of successful infection. Infected cells release free virions at a rate k, and so the rate of viral production from infected cells is ky. Uninfected target cells enter the system at rate λ and each die at a rate d, and infected cells die at a higher rate a. Free virus is cleared at a rate u. The total loss rate of each of these populations is thus ay, dx, and uv. When these terms are put together to describe the concentrations of x, y, and v over time, they give rise to the following system of differential equations:

$$\frac{dx}{dt} = \lambda - (d \cdot x) - (\beta \cdot x \cdot v)$$

$$\frac{dy}{dt} = (\beta \cdot x \cdot v) - (a \cdot y)$$

$$\frac{dv}{dt} = (k \cdot y) - (u \cdot v)$$

This system of equations must be numerically integrated for specific parameter values to understand how the infection develops over time, since it cannot be solved analytically. Examples of trajectories, and their dependence on different values of the parameters, are shown in Figure 3. However, it is possible to gain insight into some properties of the model using mathematical analysis.

The viral dynamics model has two possible equilibrium values of the variables (sets of $\{x, y, v\}$ for which the time derivatives of all three variables are zero) . We denote the

equilibrium value for the three state parameters as x^*, y^*, and v^* respectively. In one equilibrium, target cells are the only nonzero population:

$$x^* = \frac{\lambda}{d}$$
$$y^* = v^* = 0$$

In the other equilibrium, all three variables are non-zero:

$$x^* = \frac{x_0}{R_0}$$
$$y^* = (R_0 - 1)\frac{du}{\beta k}$$
$$v^* = (R_0 - 1)\frac{d}{\beta}$$

Stability analysis shows that for any given parameters set, one and only one of these equilibria is stable. Stability depends on only a single factor which is a combination of all parameters in the model, and is termed the basic reproductive rate, denoted R_0, which we define as:

$$R_0 = \frac{\beta \lambda k}{adu}$$

When $R_0 < 1$, the first "uninfected" equilibrium is always stable, implying that the infection dies out for any initial condition. It therefore also represents the steady state level of target cells before the infection has started. When $R_0 > 1$, the second "infected" equilibrium always exists ($x, y, v > 0$) and is stable, implying that the infection spreads from any initial condition to eventually reach a steady state level in the body. Beyond dictating long-term equilibrium stability, the basic reproductive ratio is actually one of the most biologically significant values that can be derived from the viral dynamics model, and is analogous to a quantity that arises in

population-level epidemic models. In this context, it can be thought of as the number of secondary individuals that a single infected individual will infect, on average, before dying or recovering, when the population is otherwise susceptible. On a smaller scale of the viral dynamics model, R_0 can be thought of as the number of secondary infected cells that arise from a single infected cell before it dies, in a body of otherwise uninfected target cells. This interpretation can be understood by examining the components of R_0. A single infected cell has an average lifespan of $1/a$, and since it produces virions at rate k, it on average produces k/a virions before dying. Each virion has an average lifespan of $1/u$, and creates additional infected cells at a rate β times the concentration of target cells. If we assume that x is at equilibrium is $\boldsymbol{\lambda}/d$ before infection occurs, then the new infection rate per virion per time is $\beta\boldsymbol{\lambda}/d$, and then over its lifespan, each virion infects an average of $\beta\boldsymbol{\lambda}/du$ new cells. Combining these terms in fact results in the mathematical expression for R_0, confirming formally the biological significance of a value that originated merely as a threshold for long term stability analysis.

In the event that $R_0 > 1$ and the infection takes off, we can approximate the initial behavior of the equations by assuming target cells are initially near their initial value ($x \sim \boldsymbol{\lambda}/d$), and so the equations become linear. In this regime, the viral load grows exponential growth rate and a rate we term r_0, where $r_0 = a\,(R_0 - 1)$. Following this initial upslope, viral load peaks, declines by some amount, and then eventually begins to stabilize towards its long term equilibrium (Figure 3). In general, higher values of $\boldsymbol{\lambda}$ and β increase the steepness of the initial upslope, and higher values of a increase the amplitude of the oscillations towards equilibrium. Higher initial viral loads from which the infection starts (v_0) shifts the curves to the right.

Based on this analysis, it is clear that if we had knowledge of these parameter values, it would tell us many important features of the infection, such as how efficiently virus infects cells, how effectively treatment must be to counteract this ability to infect, how quickly infected cells are dying, and more. Additionally, estimates to these parameters can be used to estimate composite parameters, such as the basic reproductive ratio R_0, the initial exponential growth rate r_0, the equilibrium viral load, time of peak viral load, time to reach equilibrium, viral burst size, and more. However, a major limitation is that these parameters are often difficult or even impossible to measure experimentally. Consequently, data on the time course of the infection over time must be used to infer the parameter values from the data via the process of fitting the model to the data.

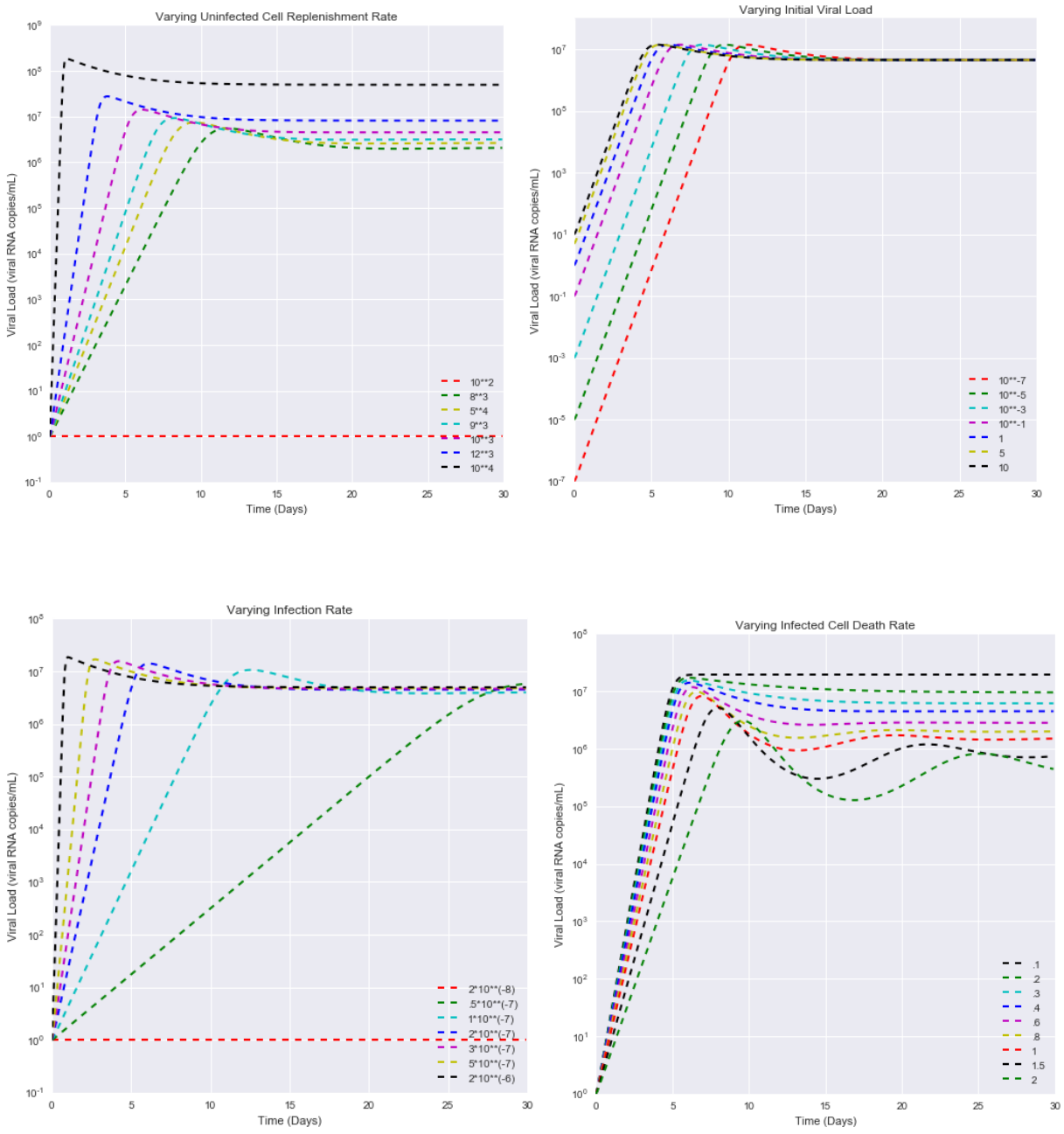


Figure 3: Dynamics of the basic viral dynamics model under different parameter values. The viral dynamics model was numerically integrated starting from the initial conditions ($x_0 = \lambda/d$, $y_0 = 0$, $v_0 = 1$) and run for 30 days. We observe how different values of λ, v_0 , β, and a impact the trajectory of viral load. When not varied, values of the parameters were $\lambda = 10^3$, $a = .4$, $\beta = 2*10^{-7}$, $k = 5*10^4$, $u = 25$, $d = .1$, and $v_0 = 1$.

Data Summary

The data available for this study are from rhesus macaques infected with SIV and treated with antiretroviral therapy (ART). SIV is often used as an animal model of HIV, as it is the closest known relative of HIV. SIV is present in non-human primates in West Africa, is non-pathogenic to humans, and shares many genetic and pathophysiological characteristics with HIV, making it an ideal model virus to use in the study of HIV (Sharp et al, 2018). The data consist of 21 rhesus macaques who were infected with SIV, treated with ART after 8 weeks of infection, and then continued on ART for 2-3 years before eventually stopping therapy. During ART, they were given either placebo treatment or multiple doses of an investigational immunotherapy, a TLR7-agonist. The Toll-like receptor 7 is a pattern recognition receptor in the innate immune system that helps recognize RNA viruses, and then stimulates the maturation of antiviral immune cells. This agonist is a small-molecule that stimulates the receptor. The data consist of longitudinal measures of viral loads (concentrations of viral RNA in the blood). They are separated into three distinct phases: *acute*, *treatment,* and *rebound*. The *acute* data measures the viral load beginning with the initial SIV infection in each individual. Viral load is measured every 7 days, from time t = 0 at the day of infection to day 56 when antiretroviral therapy began. The second stage of the data is *treatment*, which measures the viral load of the macaques when they are in the process of receiving ART (and potentially also the TLR7-agonist). Finally, in order to measure the effectivity of the immunotherapy, the monkeys were taken off of all treatment, at which point their viral load was measured every 3 days. This was known as the "rebound" stage of treatment.

These data are unique in that they capture acute infection, which is very difficult to detect in humans, as most people only experience flu-like symptoms during this time and don't realize they are HIV+ until much later on. Additionally, since allowing ART to be stopped in humans is risky due to the risk of spreading the virus to others and experiencing further deterioration of the immune system, it is rare to have good kinetic data on viral rebound in humans. Thirdly, the TLR7-agonist being tested in this study is a new investigational drug that has been proposed to potentially help cure HIV infection.

It is important to note that of the three state variables in the viral dynamics model, only one, viral load, is experimentally measured in this study. The implication of this is that any method used to fit this data to the model must be able to identify all parameter values for the system of equations, with viral load data over time as the only available information. Another aspect of the data that is particularly noteworthy is that due to lack of sensitivity of the equipment used to measure the viral load in the individual monkeys, any viral load below the detection limit of 50 viral RNA copies/mL is undetectable. We refer to any viral load measurement observed to be below this detection limit of 50 copies/mL as censored data. The only information that these censored viral load data points carry is assurance that the viral load at that time was some value below 50 copies/mL. There is no way to identify its exact value. These two aspects of the data present particular challenges to a model-fitting approach to estimate the parameters of the viral dynamics model.

Fitting method

Our goal is to estimate the parameters of the viral dynamics model introduced above using the data from an animal model of HIV (SIV infected rhesus macaques). Conceptually, model fitting involves identifying parameter values that minimize the distance between the observed data and the model (solutions obtained by numerically integrating the system of differential equations over time for these parameter values). To compare the model to data, I first considered as an objective function the sum of the residuals, or the difference between the data and the model. Since viral load values change over many orders of magnitude during infection (Figure 3), I implemented this on a log (base *e*) scale:

$$SSR(\theta) = \sum_{i=1}^{n} (log(v(t_i)) - log(\hat{v}(t_i|\theta))$$
$$= \sum_{i=1}^{n} log(\frac{v(t_i}{\hat{v}(t_i|\theta)})$$

Where $v(t_i)$ is the observed viral load at time t_i, $\widehat{v}\,(t_i|\theta)$ is the viral load at time t_i simulated by the model under a set of parameter values θ. This difference is summed over all n data points. The package I utilized for this process is called lmfit, which can be installed in Python for fitting complex functions with non-linear least squares fitting algorithms (Newville et al 2014). The general structure of the lmfit package allows for defining a model for your data, an objective function, a set of plausible ranges for all parameters being estimated and an initial guess at which to start for each, and a minimizer of your choosing. The minimizer acts on the objective function, and does so by testing out values within the range of plausible values provided for each parameter to minimize the objective function with respect to the data that it is

fitting. The set of parameters that it converges on to minimize the objective function are the best fit parameters that enable the model to most closely resemble the data. Built into lmfit is the flexibility to select a minimization algorithm, with the default being the Levenberg-Marquardt least squares method, a standard method for performing nonlinear least squares problems by minimizing the sums of squared error between an array of data points and a model array.

I used this method to fit the viral dynamics model to data during acute infection, during ART, and during rebound following ART-cessation. For acute infection (*acute* phase), because the animals were infected by injecting a small amount of viral particles, we set $v(0) = v_0$ (an extra parameter of the model that needs to be fit), and we assume that infected cells are at their equilibrium concentration ($x_0 = \lambda/d$). We assume there are no infected cells at time zero ($y_0 = 0$). During antiretroviral therapy (*treatment* phase), we assume $\beta = 0$, as the drugs are extremely effective at blocking viral infection. Under this assumption, the model becomes linear. Viral load decays bi-exponentially, but in the limit where viral production is large and viral clearance is high compared to other parameters (the case for HIV and SIV, further clarified in the next section), it decays exponentially with rate a (so only this parameter can be identified then). During the rebound phase, the infection restarts when a latently infected cell "wakes up" and starts producing virus again. We assume that the length of the ART phase is long enough so that target cells have completely recovered to their pre-infection levels). In this case, the initial condition is ($x(0) = \lambda/d$, $y(0) = y_0$, and $v(0) = 0$).

Results

Sensitivity Testing

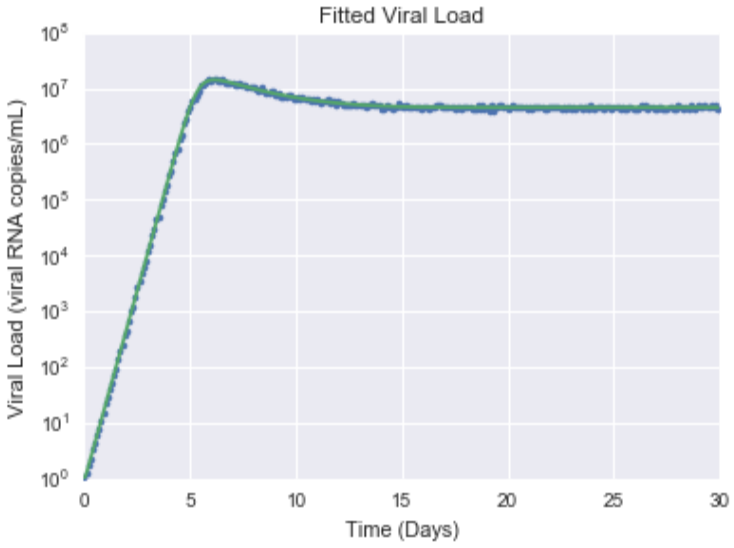
Evaluation of the fitting algorithm on simulated data

Before utilizing this least squares curve fitting the recover the parameter values from the experimental viral load data, I first had to perform sensitivity testing to confirm that this method can in fact identify the true underlying parameter values. I did this by generating data from a particular set of parameter values, performing the fitting procedure on this generated data, and observing whether or not the true best fit parameters were recovered. By first checking the accuracy and consistency of the fitting method when the underlying parameter values were known, this provides the experimental basis necessary to apply the method on real data and have confidence in the parameter values it recovers. In order to properly perform sensitivity testing to establish confidence in the robustness of the procedure and highlight its limitations, it was necessary to mimic the conditions of, and challenges posed by, fitting the real data.

Prior research has made the point that because the kinetics of changes in viral load are much faster than those of uninfected or infected target cells, virus very quickly (timescale of ~ 1 hour) reaches a quasi-steady state level with respect to infected cells, meaning that viral loads are essentially proportional to infected cell counts ($v \sim (k/u)*y$) (Nowak & May, 2000). As a result, when all state parameters are observed, it is impossible to uniquely identify both viral burst rate k and viral clearance rate u from the data; only the ratio is identifiable. Luckily, the viral clearance

rate is one parameter that is possible to measure experimentally, and it has been estimated to be u = 23 day^{-1} (Ramratnam et al 1999). Because of this, throughout the rest of the paper whether fitting simulated or actual data, I fix u at this value.

To provide the simplest check of the method, I first tested the curve fitting algorithm on simulated data in which all three state parameters, uninfected cells, infected cells, and free viral load, were observed. Additionally, the data were essentially continuous with data points every tenth of a day, and minimal noise such that data points deviated from their true value following a Log-normal distribution with mean 0 and standard deviation .05. The parameters that generated the simulated data were set to have true values $\lambda = 10^3$, $\beta = 2 * 10^{-7}$, $k = 5 * 10^4$, $d = .1$, $a = .4$, $u = 25$, and $v_0 = 1$, though additional values were ultimately tested in different stages of the process of sensitivity testing to check the robustness of the fitting algorithm on viral load data with vastly different dynamics, as would be expected in real data. It should be noted that in this simulation I am acting as though all three state parameters were observed. Therefore the objective function that is being minimized in the process of the nonlinear least squares model fitting is not simply the difference between the model and data for viral load, i.e. $\hat{v}(t_i|\theta) - v(t_i)$, but is instead finding the set of parameter values θ that minimize the sum of this distance squared combined with the sums of squared residual for both x and y, in addition. The results of this fit can be seen in Figure 4, highlighting both a visual of the model overlaying the data, as well as values quantifying the accuracy of the fit.



Model AIC: -5104.207		
Parameter true value	Best Fit Value	Uncertainty
$V_0 = 1$	.994	.27%
$\beta = 2*10^{-7}$	$2*10^{-7}$	.38%
$\lambda = 10^3$	986.68	.21%
$d = 0.1$	0.0984	.42%
$a = 0.4$	.3965	.23%
$k = 5*10^4$	$5.02*10^4$	.39%

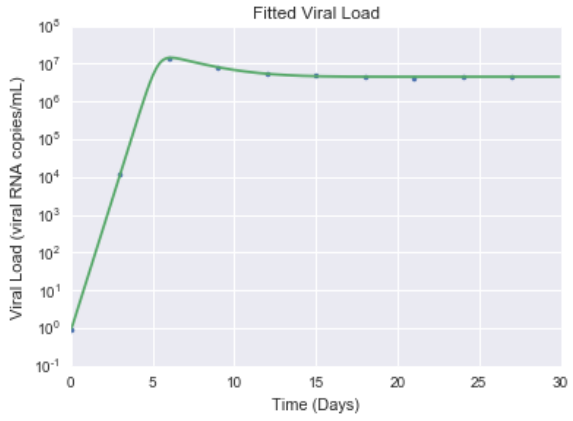
Figure 4: Summary of preliminary method check a.) Visualization of best fit viral load solution b.) Overall model accuracy as measured by AIC, comparison of each parameter best fit value to the true value, as well as uncertainty in parameter estimation.

The results show that under these ideal circumstances, the model is able to simultaneously estimate all six unknown parameter values with a high degree of certainty. Uncertainty is calculated from the 1 standard deviation quantile about the medians of the probability distribution underlying the parameter optimization within lmfit. This measure of uncertainty is automatically calculated with each iteration of the lmfit fitting procedure. The low AIC (which is a probabilistic measure of fit accuracy in which a lower value signifies a better fit of the model to the data), proximity of best values to the true values of the parameters that generated the data, and low uncertainty are indication of a very successful fit. With the baseline confirmed, I then varied the frequency of observations, the amount of noise in the data, the number of parameters being estimated at any one given time, the number of state parameters observed, and the presence or lack thereof of censored data for various sets of parameter values

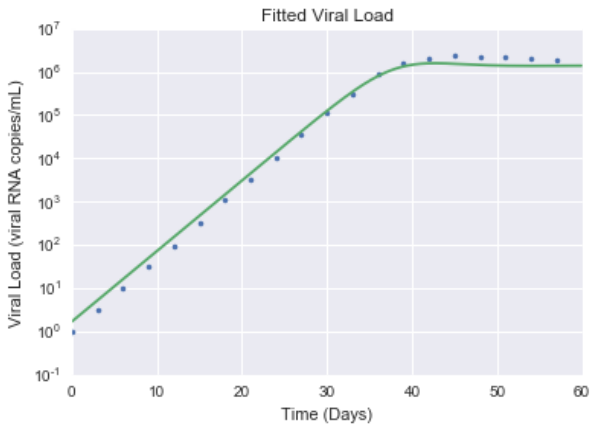
to confirm the robustness of the method to deviations from the ideal that are likely to be encountered in real experimental viral load data.

Robustness of fitting algorithm to sparsely sampled time points

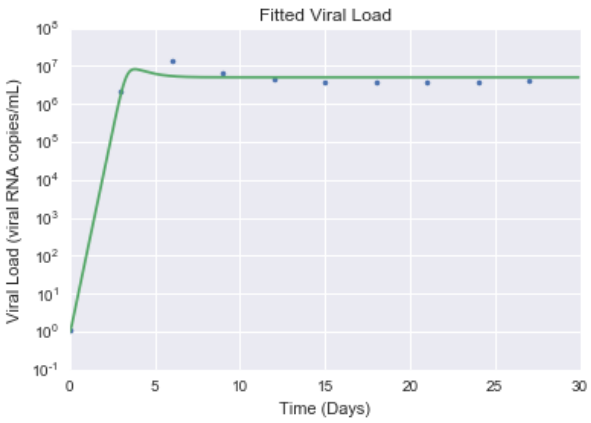
The fitting procedure needs to be able to recover the true parameter values when the data is sparsely rather than essentially continuously sampled. To account for this, I adjusted the frequency of time points in the simulated data to every three days, which is more indicative of the actual frequency of observations in the available data for this study. Three sets of parameter values with different kinetics were used to ensure ability of the method to fit the model to data that don't resemble each other.



Model AIC: -166.4		
Parameter true value	Best Fit Value	Uncertainty
$V_0 = 1$	1.04	4.05%
$\beta = 2*10^{-7}$	$2.06*10^{-7}$	2.76%
$\lambda = 10^3$	1021.6	2.93%
$d = 0.1$	0.104	4.15 %
$a = 0.4$	.404	2.07%
$k = 5*10^4$	$2.94*10^4$	2.51%



Model AIC: -85.99		
Parameter true value	Best Fit Value	Uncertainty
$V_0 = 1$	1.7	9.27%
$\beta = 1*10^{-7}$	$8.9*10^{-8}$	14.36%
$\lambda = 10^3$	1899.4	7.57%
$d = 0.2$	0.31	2.64%
$a = 0.4$	1.00	1.24%
$k = 4*10^4$	$6.4*10^4$	13.85%



Model AIC: -34.1		
Parameter true value	Best Fit Value	Uncertainty
$V_0 = 1$	.99	37.15%
$\beta = 1*10^{-7}$	$5.8*10^{-7}$	21.28%
$\lambda = 8*10^2$	$3*10^3$	10.31%
$d = 0.05$	0.48	.87%
$a = 0.4$	1.00	13.99%
$k = 4*10^4$	$4.8*10^4$	22.59%

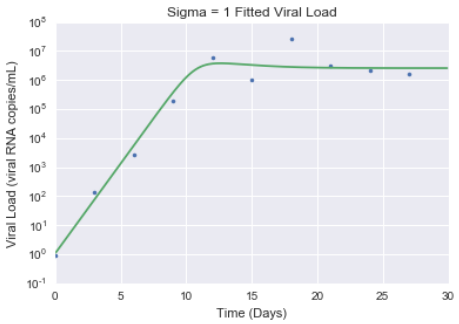
Figure 5: Summary of method check on sparse data. Plots of different solutions for varying values of of the parameters and accompanying summary statistics. a) Baseline parameter values: $\boldsymbol{\lambda} = 10^3$, $\beta = 2 * 10^{-7}$, $k = 5 * 10^4$, $d = .1$, $a = 0.4$, $u = 25$, and $v_0 = 1$. b) Altered to reflect kinetics of lower R_0(~1.5): $\boldsymbol{\lambda} = 10^3$, $\beta = 1 * 10^{-7}$, $k = 4 * 10^4$, $d = .2$, $a = 0.4$, $u = 25$, and $v_0 = 1$ c) Altered to reflect kinetics of a>>d: $\boldsymbol{\lambda} = 8*10^2$, $\beta = 2 * 10^{-7}$, $k = 5 * 10^4$, $d = .05$, $a = 0.4$, $u = 25$, and $v_0 = 1$

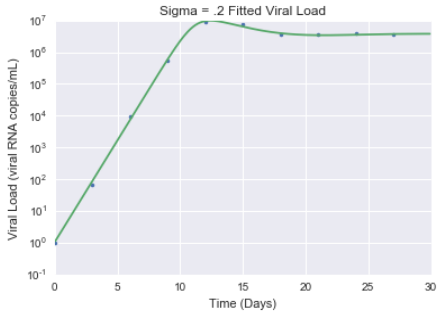
These results show that the method was confirmed to be robust to varying frequency of observations. There was certainly a decrease in the fit accuracy as indicated by distance of the best fit values from the true values, the AIC, as well as the uncertainty in the parameter estimations, as is indicated by the higher values for AIC and parameter estimation uncertainty for all three sets of simulated data. The ability of the algorithm to recover parameter estimates within a plausible range of the underlying values shows that the fit is still quite successful under these conditions. This confirms that frequency of observations in the real experimental viral load data will not be preventative of successful recovery of the true parameter values. One limitation that this stage of testing did highlight is that sparsity of data can potentially pose a problem towards estimating true parameter values with a high degree of precision, particularly when there are few time points early on, as seen in Figure 5c. It appears that in the case of a steep initial growth rate and infrequent sampling such that few observations are available leading up to the peak, there is difficulty identifying the early exponential growth rate as well as the peak value. This does not,

however, appear to impact the ability of the model to find the set point equilibrium viral load in the long run. This is seen both in the fact that the fit depicted in Figure 5c visually isn't exactly picking up on the correct pattern in the data, and several of the parameter values are quite far from their true value (such as *d* in Figure 5c). These challenges posed by minimal data on the early exponential viral growth rate as an artifact of sparser data are important to recognize, but overall do not appear to be inhibitory to fitting data that were sampled at comparable frequency.

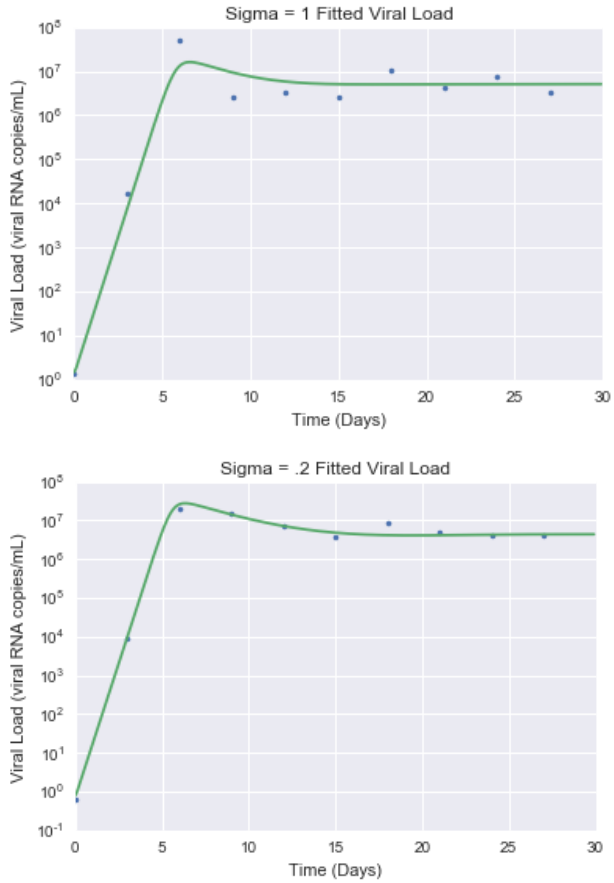
Robustness of fitting algorithm to sampling error in viral load values

In order to mimic the noise inherent in real data, the next step in sensitivity testing the fitting method was to confirm its robustness to deviation from the true underlying value generated by the model. Recall that the initial value for the standard deviation of the Lognormal distribution that the error term was being drawn from was .05. The results of increasing the amount of error to .2 and then 1, with all else held equal from the earlier described conditions in the previous section, are shown below.





$\sigma = .2$			$\sigma = 1$		
Model AIC: - 80.857			**Model AIC: 12.230**		
Parameter	Best Fit Value	Uncertainty	Parameter	Best Fit Value	Uncertainty
$V_0 = 1$	1.106	7.93%	$V_0 = 1$	1.11	70.11%
$\beta = 10^{-7}$	$9.9*10^{-8}$	10.74%	$\beta = 10^{-7}$	$3.15*10^{-7}$	61.05%
$\lambda = 10^3$	$1.2*10^3$	12.68%	$\lambda = 10^3$	$1.1*10^4$	83.32%
$d = 0.1$	.12	15.33%	$d = 0.1$	.17	108.45%
$a = .4$	.52	5.98%	$a = .4$	.33	57.19%
$k = 5*10^4$	$5.5*10^4$	7.6%	$k = 5*10^4$	$2.3*10^4$	29.24%



σ = .2			σ=1		
Model AIC: - 73.495			**Model AIC: 10.671**		
Parameter	Best Fit Value	Uncertainty	Parameter	Best Fit Value	Uncertainty
$V_0 = 1$	.79	18.9%	$V_0 = 1$	1.5	77.75%
$\beta = 10^{-7}$	$1*10^{-7}$	11.93%	$\beta = 10^{-7}$	$8.12*10^{-8}$	53.56%
$\lambda = 10^3$	912.1	10.48%	$\lambda = 10^3$	$1.4*10^3$	55.74%
$d = 0.05$	.045	14.1%	$d = 0.05$	.11	76.8%
$a = .4$	..38	6.51%	$a = .4$	.47	39.38%
$k = 5*10^4$	$4.9*10^4$	11.74%	$k = 5*10^4$	$4.8*10^4$	48.67%

Figure 6: Results of robustness check to sampling error. Two different sets of model parameters are used to assess the ability of the model to fit the data as sampling error increases, and highlight ability of the method to accurately fit the model to data with non trivial amounts of error.

As anticipated, increasing the amount of noise in the data makes it more difficult for the fitting model to converge upon the optimal parameter values. This is reflected in the increase in the uncertainty in the estimation of all parameters as noise level was increased from a standard deviation of .2 to 1 on the log scale. This increase in parameter uncertainty was accompanied by an overall decrease in the accuracy of the fit. With data deviating from its true value with a standard deviation of 1 on the log scale, the optimal parameters that the fitting algorithm converged upon resulted in AIC values of 12.23 and 10.671 respectively, much higher than the values of -80.857 and -73.495 when the error was sampled from a distribution with standard

deviation .2. This highlights that data that inherently have higher error lead to more uncertainty in estimation, as would be expected intuitively. This uncertainty, however, is not inhibitory. It is clear that even in the case of a high degree of error in the data, we are still able to recover parameter estimates in aggregate that are incredibly close to their true underlying values.

Identifiability of the model with unobserved state variables

I next tested the fitting method with only the viral load data observable. This means that the fitting algorithm had to be altered slightly such that the residuals being minimized through the process of non-linear least squares fitting were only based on viral load, which is the only data that is experimentally available. Previous work has shown that when only viral load is measured, the parameters k (viral burst rate) and λ (target cell input rate) are not mutually identifiable, meaning that only product of them can be identified (Borducci et al, 2016). Intuitively, this is because the target cell population is proportional with λ, while the ratio between infected cells and virus is proportional to k. Any particular viral load trajectory could always be caused by either a high population of target cells and a low viral production rate, or, a smaller population of target cells but a lower viral production rate. Consequently, it is best to fix one of these parameters and estimate the value of the other from data. Because there are reasonable estimates of the viral burst rate, we chose to fix $k=5 * 10^{-4}$ (virions/cell/)(day^{-1}) (Chen et al 2007). This is in addition to the previously fixed value for the clearance rate of free virus u at 23 (day^{-1}).

When testing the fitting algorithm within the previously noted conditions (k, and u fixed, sparse data, non trivial amount of noise) now with only viral load data observed, it became clear

that due to the underlying stochasticity of the algorithm, the lmfit minimizer was sometimes converging on a suboptimal solution. Additionally, there was a high degree of variability in the solution that the algorithm was converging on, meaning that it was returning a different fit each time, with some fitting the data quite well, and others converging on a suboptimal solution. Thus there was no longer evidence to trust the output of a single iteration of the fitting algorithm. This problem was overcome by taking an iterative optimization approach to the fitting process. This means that instead of just running the least squares minimization one time through, I instead iterated through the process 100 times, with the initial guess for each parameter value coming from a uniform distribution constructed conservatively such that it contained a wide range of plausible values so as not to reflect too much a priori knowledge about the true parameter values.

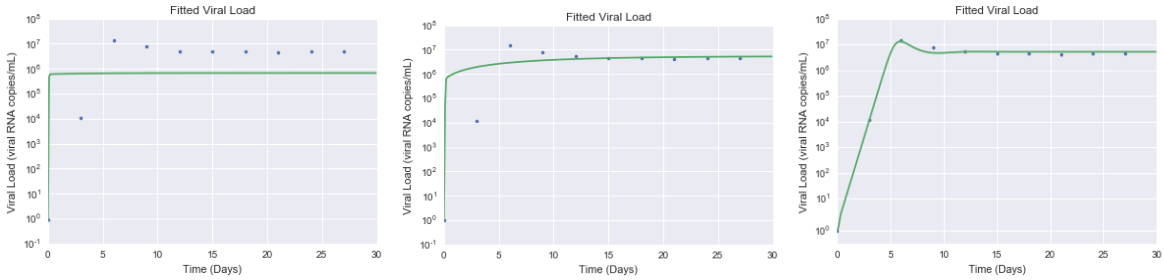


Figure 7: Contrasting single solution with iterative optimization. Side by side comparison of two single iteration best fit solutions for data generated from underlying parameter values $\lambda = 10^3$, $\beta = 2 * 10^{-7}$, $k = 5 * 10^4$, $d = .1$, $a = 0.4$, $u = 25$, and $v_0 = 1$ (left and center) and iterative optimization to find best fit solution with n = 100 iterations (right).

This method of iterating over the fitting algorithm repeatedly allows for isolating the parameter values that minimize the sums of square residuals across all 100 best fit solutions. Varying the initial guess was necessary for ultimate convergence on the true parameter values, as I found that just iteration without varying the initial guess for the minimization did not lead to conversion towards the true parameter values. For example, for the same set of parameters used to simulate the data depicted in Figure 7, with error added to the viral load data with standard

deviation .3 on the log base e scale, when the fitting algorithm was optimized over 100 iterations when the initial guess was the same every time, the AIC for the best fit solution was 39.4. In stark contrast to this, when the same process was done on identical data with the only difference being the initial guess for each parameter being estimated now being sampled randomly from a uniform distribution, the resulting AIC associated with the optimal solution across the 100 iterations was -58.5. This confirmed the benefit to varying the initial guess for each parameter in order to improve parameter estimates and converge to an optimal solution. Thus with confirmation of the ability of iterations through lmfit with a varying initial guess with each iteration, and an understanding of the limitations of the fitting algorithm in terms of which parameters can be simultaneously estimated, I was confident in the ability of this approach to produce reliable best fit parameter estimates from real viral load data.

Fitting on experimental data

Censored viral load values and maximum likelihood expectation

Given the confidence in the fitting procedure developed through sensitivity testing, the first method attempted on the actual data was exactly inline with the procedure used to recover parameter estimates from the simulated data. To reiterate, this involved fixing viral death rate, u, and rate at which free virus is produced by infected cells, k, at their experimentally estimated values, and iterating through the lmfit least squares fitting algorithm 100 times with initial parameter guesses for each of the 5 parameters being estimated being drawn from a uniform distribution in order to find the parameter estimates that minimized the sums of squared error

between the model and the data. When this procedure was performed on the rebound data, it gave rise to several problems. The first was in determining a range of plausible values to set the bounds both of the uniform distribution that determined the initial guess for each parameter, as well as for the argument that set the minimum and maximum value that each parameter could take on. In order to ensure that these ranges weren't excluding the true values of the parameters, I used wide ranges informed by other studies using similar data or models (personal communication from Alison Hill and Jeff Gerold). For the five parameters being estimated, v_0 and a could take on any value between 0 and 10, λ could take on any value between .01 and 10^3, d could take on any value between 0 and 50, and β could take on any value between 0 and .01. The initial guesses for each parameter were drawn from a uniform distribution that was much narrower in order to make it easier for the fitting algorithm to recover the optimal solution.

By virtue of the detection limit of the data, almost every individual in the rebound phase had undetectable viral load values (< 50 virions/mL, "censored data") up through day 8 after being taken off ART. To account for this, I initially took only the last undetectable value to be exactly 50 copies/mL (ignoring earlier time points), plus all subsequent data points above the threshold, to use as input for the fitting algorithm. The hope was that doing so would provide enough information early on to approximate the early growth in viral load, as well as fit to the eventual peak and return to set point equilibrium. In using just the last value of 50 copies/mL onward, the best fit parameters that minimized the residuals across the 100 iterations (and even when iterations were increased to 500) was still producing a suboptimal solution, with particular trouble fitting the initial upslope of the data. Examples of the difficulty fitting the initial viral growth on all the censored data are illustrated in Figure 8.

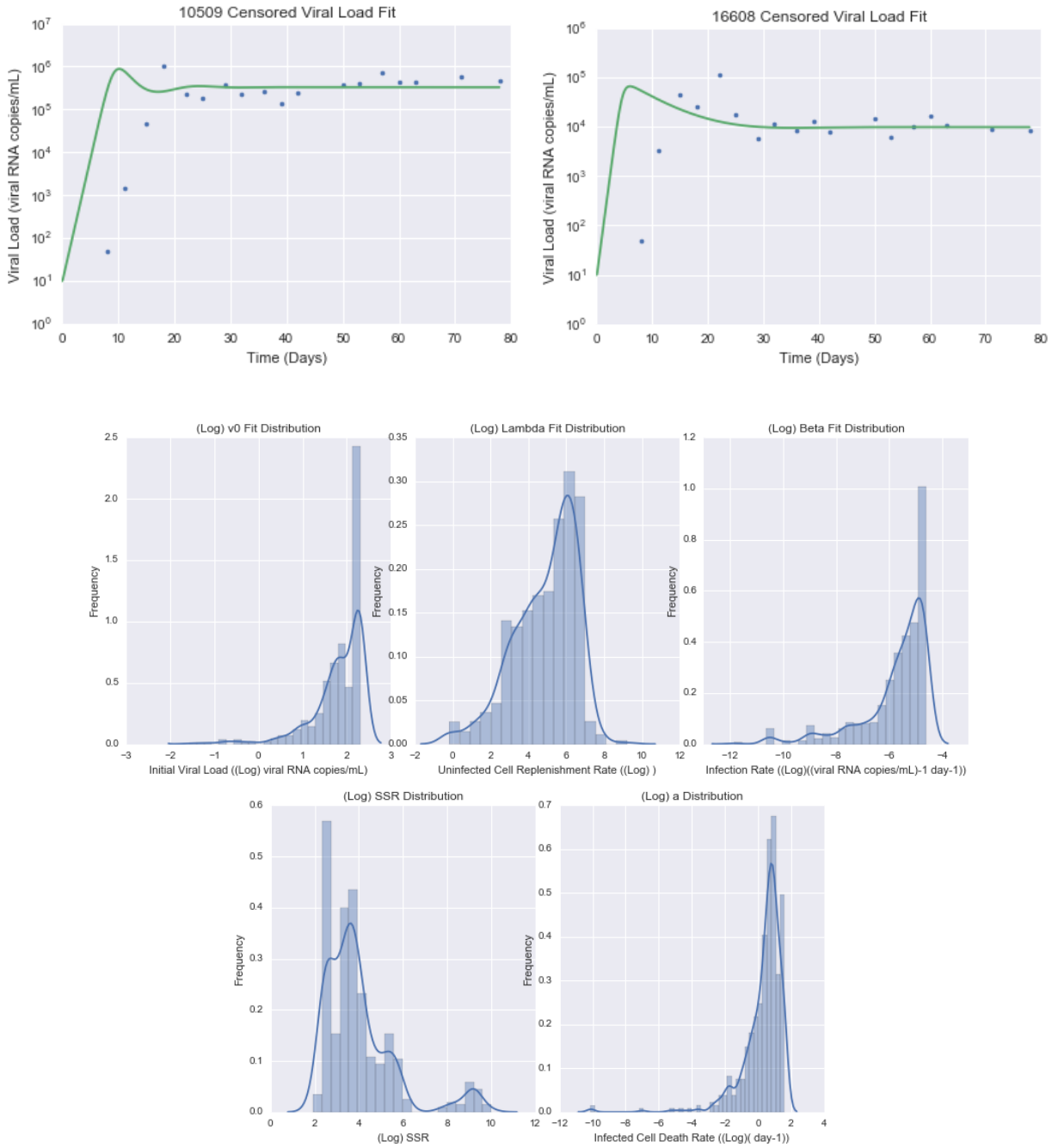


Figure 8: Fitting subjects 166-08 and 105-09 on censored data from last censored value on. Viral load values < 50 copies/mL cannot be quantified and are considered "censored" data. All subjects have censored data at early time points. Both solution plots and the parameter histogram highlight the inability of the algorithm to fit the data successfully, particularly on the initial upslope. There is a high degree of uncertainty in estimation of the initial viral load, v_0, as well as in the target cell replenishment rate λ.

An important takeaway from this analysis is that v_0 estimates are approaching the upper limit of 10 copies/mL, which is far higher than initial viral load should be following ART

treatment. This inability to identify a reasonable estimate for v_0 given the nature of the data censored below values of 50 particles/mL highlighted a limitation of the fitting algorithm in its current form: there is too much uncertainty to approximate the initial viral load slope. Where this approach is lacking is in its inability to draw information from the values of 50 copies/mL in the viral load data, which are informative in that they let us know *at most* what the data could have been. While this certainly is not as helpful as precise measurements would be towards identifying parameter values that allow the model to accurately fit the data, the censored data can still be informative in the fitting process through taking a probabilistic approach. In order to make use of the information the data was providing in its entirety and take into account these maximum thresholds of the censored data, the fitting algorithm had to be adjusted slightly by implementing a maximum likelihood approach to the data fitting.

Maximum likelihood estimation overview and implementation

In order to take advantage of all the information available in the data, rather than excluding the data points from below the threshold, I took a probabilistic approach to the data fitting. Recall that previously in my initial approach to fitting the data I was looking to minimize the residuals, or the distance between each point and the model on the log (base e) scale. When taking a probabilistic approach, we are instead trying to maximize the likelihood of seeing the data observed given a certain model. As a parallel to the process of minimizing the distance between data and model, in the probabilistic regime we instead find the parameters such that the probability of observing the data is maximized. Because the data are coming from an unknown distribution without a closed form, we apply the maximum likelihood approach to the error about

the true viral load. This is done by assuming that the observed viral load follows a normal distribution on the log scale about the theoretical mean viral load value at a given time point, such that the observed viral load data, $\hat{v}(t) \sim$ Lognormal(μ, σ^2). Therefore, the probability of observing any one data point, $v(t_i)$ given a model with a specific set of parameters θ, notated as $\hat{v}(t_i|\theta)$, can be written using the probability density function (PDF) of a lognormal distribution as:

$$\begin{aligned} P(\text{Data}_i \mid \text{Model}) &= P(v(t_i) \mid \theta) \\ &= f(\log(v(t_i)) \mid \log(\hat{v}(t_i|\theta)), \sigma^2) \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\log(v(t_i)) - \log(\hat{v}(t_i|\theta)))}{2\sigma^2}} \end{aligned}$$

Assuming independence of observations and therefore independence of error that the observed viral load data is sampled from, such $\varepsilon_i \sim Lognormal(0, \sigma^2)$ for $i = 1...n$, the joint probability of observing the data collectively is simply the product of the probability of observing any one data point, across all n observations. We call this joint probability of *all* the data given the model the likelihood function, and notate it as:

$$\begin{aligned} \mathcal{L} &\equiv P(\text{Data} \mid \text{Model}) \\ \mathcal{L}(\{v(t_i)\} \mid \theta) &\equiv P(\{v(t_i)\} \mid \theta) \\ &= \prod_i f(\log(v(t_i)) \mid \log(\hat{v}(t_i|\theta)), \sigma^2) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\log(v(t_i)) - \log(\hat{v}(t_i|\theta)))^2}{2\sigma^2}} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\sum_i (\log(v(t_i)) - \log(\hat{v}(t_i|\theta)))^2}{2\sigma^2}} \end{aligned}$$

For computational ease of dealing with sums rather than products, this is often transformed to the log (base e) scale. The log likelihood is thus as follows:

$$\mathcal{LL} \equiv \ln(\mathcal{L})$$

$$\mathcal{LL}(\{v(t_i)\} \mid \theta) = -\left(\ln(\sigma\sqrt{2\pi}) + \frac{1}{2\sigma^2}\sum_i (\log(v(t_i)) - \log(\hat{v}(t_i \mid \theta)))^2\right)$$

In this form, there is a clear parallel between the maximization of the log likelihood function and minimization of the sums of residuals, because of the monotonicity of the log function in θ. Therefore as discussed thus far, the two approaches to devising an objective function, minimization of residuals and maximization of likelihood, are identical. The advantage, however, of using a probabilistic to fitting the data becomes clear when the censored data, or data below the detection limit is taken into account. This is done by utilizing the cumulative density distribution (CDF) of the same lognormal distribution that we are assuming the error is coming from. Thus for viral load observations below the detection limit, $v(t_j) < 50$, the probability of observing any individual point t_j below the detection limit is written as:

$$\begin{aligned}\mathcal{L}(d \mid \theta) &= P(v(t_j) < d \mid \theta) \\ &= F(\log(d) \mid \log(\hat{v}(t_i|\theta)), \sigma^2) \\ &= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{(\log(d) - \log(\hat{v}(t_i|\theta)))}{\sigma\sqrt{2}}\right)\right)\end{aligned}$$

Where erf is the error function, $\operatorname{erf} = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dx$.

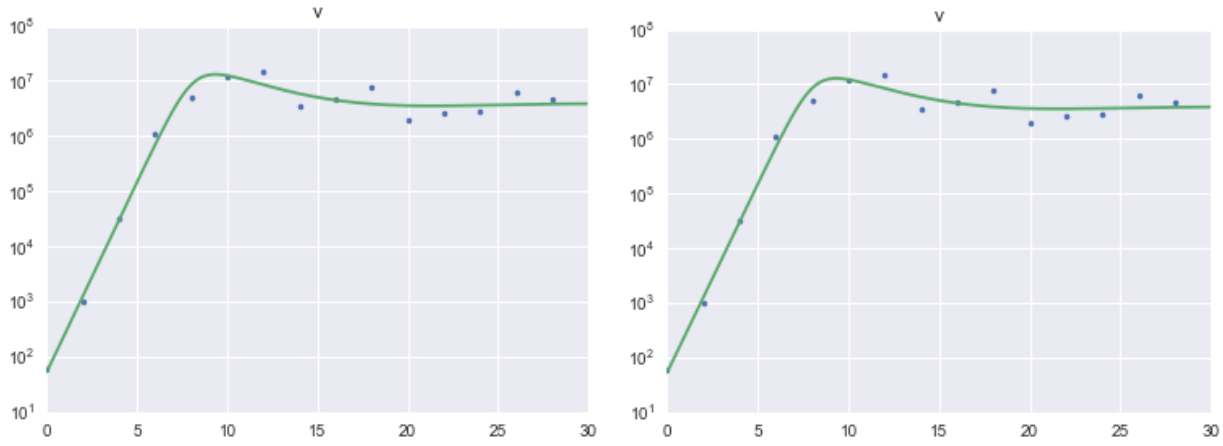
Applying the same assumptions of independence that were utilized for the data above the detection limit, the likelihood then of observing *all* the data below the detection limit given a particular set of parameters is the joint probability of each individual point, written as:

$$\mathcal{L}(\{v(t_i)\} \mid \theta) = \prod_j f(\log(v(t_i)) \mid \log(\hat{v}(t_i|\theta)), \sigma^2) \prod_k F(\log(d) \mid \log(\hat{v}(t_i|\theta)), \sigma^2)$$

Thus for a set of data points {i} such that subset {j} are above the detection limit and subset {k} are below the detection limit, the collective log likelihood function for the data can be written as:

$$\mathcal{LL}(\{v(t_i)\} \mid \theta) = \sum_j \log\left(f(\log(v(t_i)) \mid \log(\hat{v}(t_i|\theta)), \sigma^2)\right) + \sum_k \log\left(F(\log(d) \mid \log(\hat{v}(t_i|\theta)), \sigma^2)\right)$$

In order to confirm the ability of the likelihood-based fitting algorithm to fit the model to the data, I compared the solution produced by this fitting method to that of the already sensitivity-tested least squares model. This can be done because in the case of non-censored data, the two objective functions are functionally the same, and thus the least squares and MLE solutions should be identical. Thus in order to test the MLE approach to fitting the viral load data, this solution was compared to the previously tested solution using the least squares objective function under conditions without censored data. Figure 9 highlights the similarity of the two solutions, and therefore confirmation of the MLE approach to accurately fit simulated data.



Method	$\widehat{\lambda}$	$\widehat{a}$	$\widehat{\beta}$	$\widehat{d}$	$\widehat{v_0}$
OLS	832.27	.38	$8.7*10^{-8}$	.07	56.8
MLE	800.02	.37	$9*10^{-8}$	.07	56.7

Figure 9: a) Side by side residual minimization (left) vs maximum likelihood estimation (right), b) Table of best fit parameters for both fitting methods. The data were simulated from true parameter values: $\lambda = 10^3$, $a = .4$, $\beta = 1*10^{-7}$, $k = 5*10^4$, $u = 23$, $d = .1$, and $v_0 = 50$. It should be noted that although

this is *not* a realistic value for v_0, this had to increase so as to ensure that none of the data would equal 50, and thus the CDF term in the objective function would go unused.

This provided confirmation of the accuracy of the MLE fitting algorithm under these circumstances. I thus adjusted my fitting algorithm by now using the negative log likelihood of the above form as the objective function input to the minimizer, rather than simply the residual scalar difference between the model and the data. It should be noted that because the objective function is now returning a scalar value rather than an array of values, the minimization method needs to be adjusted accordingly. Instead of the previous Levenberg-Marquardt method, I now used Powell's method, a direct search method that is preferable because it does not require calculation of the first derivative of the function being optimized. For optimization problems such as the one we are dealing with that are highly non linear and don't have very many independent variables, previous studies have found Powell's method to be one of the most consistently successful. Additionally, comparison between other potential optimization candidates available within the lmfit package, including Conjugate-Gradient, nelder-mead, and Newton-Conjugate-Gradient showed that all methods available produced similar fit results, with Powell having a marginally faster run time. (Box, 1966)

Lack of practical identifiability of all parameters in rebound data

This maximum likelihood fitting algorithm was run on the rebound data first, with *k* fixed at $5*10^4$, *u* fixed at 23, and all six other parameters (five from the differential equations themselves, one to estimate the variance of the error term about the true mean viral load). After running each fit with 500 iterations of the maximum likelihood–based fitting method and

selecting the set of parameters across all 500 solutions that minimize the negative log likelihood, I returned a plot of the best fit solutions to visually check success of the fit, the values of the best fit parameters, and the minimum negative log likelihood value. Preliminary visual checks of plots overlaying the data and the solution comprised of the best fit parameters from the model fitting showed a successful fit to the data. Several examples of implementation of this iteration of model fitting are shown in Figure 10.

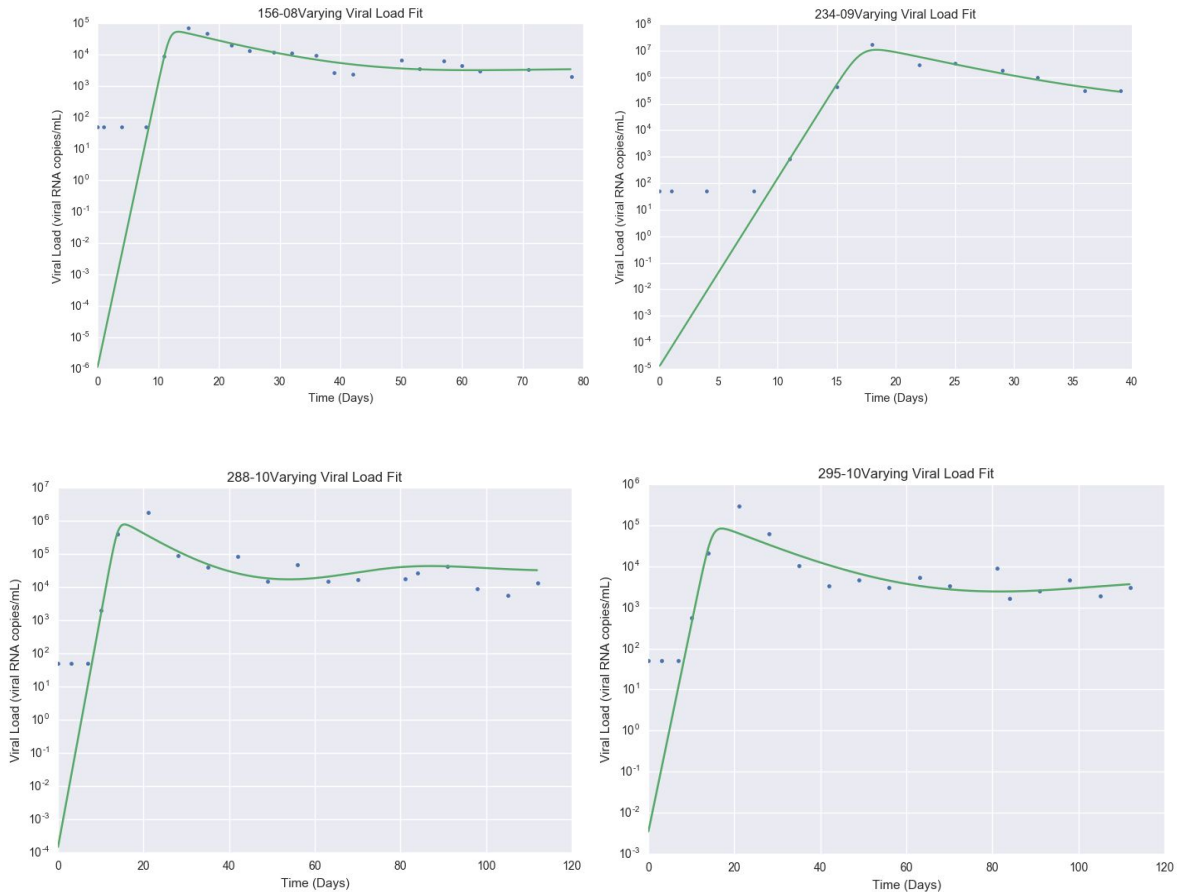


Figure 10: Sample solutions of MLE data fitting (subjects 156-08, 234-09, 288-10, 295-10). Fixed parameters were $k = 5*10^4$, $u = 23$, number of iterations over different initial conditions = 500.

While visual inspection of plots overlaying the best fit solution on the rebound data for all 19 individuals provided an initial confirmation of the ability of the model to successfully fit

the viral load data, it is also crucial to observe the uncertainty in estimation and therefore the identifiability of the parameters being estimated. This is indicative of the extent to which we are confident in the parameter estimates that the fitting algorithm is converging on, as a high degree of uncertainty in a parameter estimate or a high degree of correlation with another parameter estimate can provide a wide range of parameter values over which the fitting algorithm is equally likely to converge. This is problematic if we hope to be certain of our final parameter estimates. Thus to check for the above issues of uncertainty and correlation, I viewed histograms of the distribution of best fit parameters across the 500 iterations, and looked at heatmaps to observe any correlation between parameters being estimated. Figure 11 shows an example from a sample subject, 156-08, to highlight the correlation trends were more or less consistent across individuals.

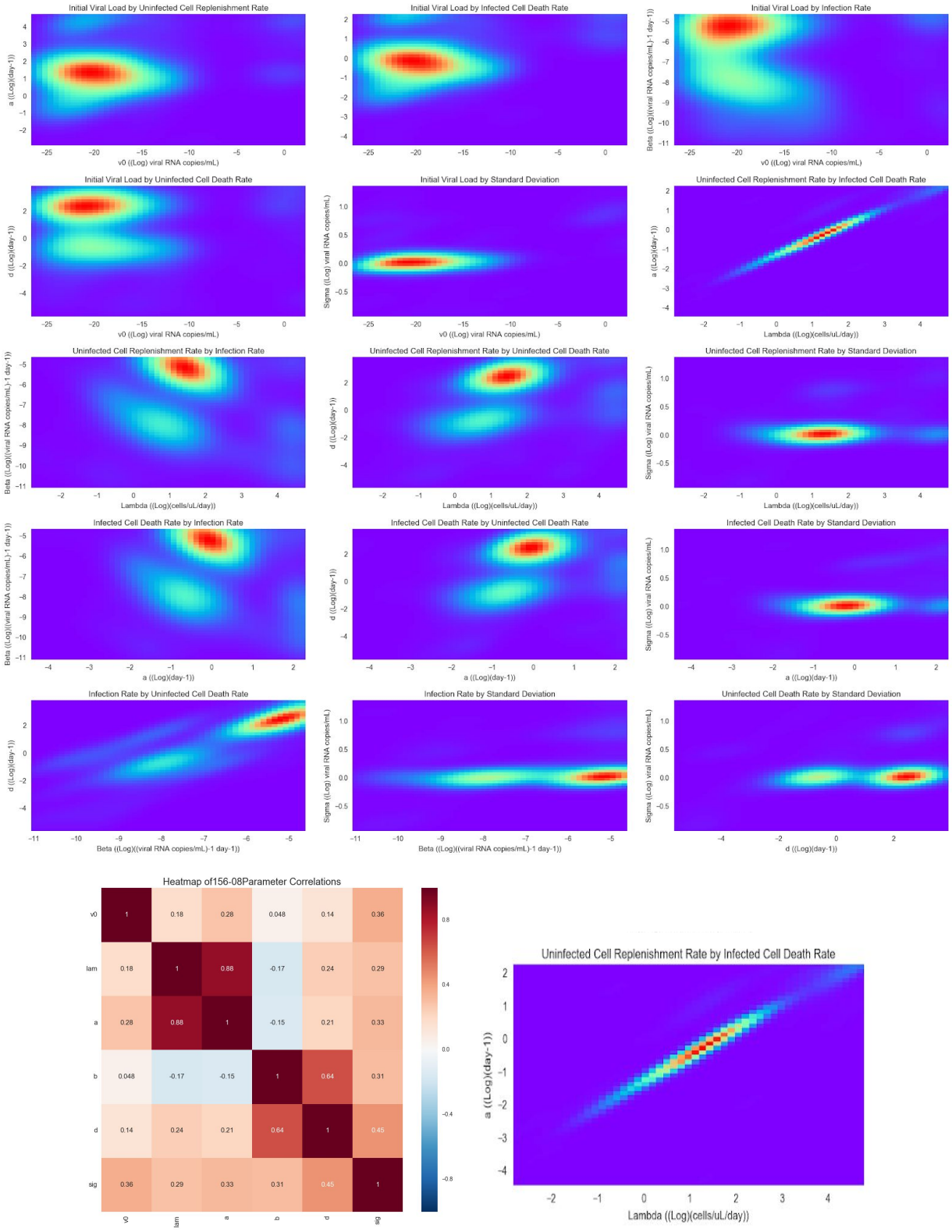


Figure 11: Correlation and uncertainty visualizations for subject 156-08 (only *k* and *u* fixed). For each subject, the maximum likelihood fitting algorithm was run 500 times with different initial

conditions, and each time the parameters that were found by lmfit to maximize the likelihood were returned. a) On each density plot the x-axis is one parameter and the y-axis is another, and the color represents the frequency of a particular combination of these parameters in an optimal parameter set returned by lmfit. b) The Pearson correlation coefficient calculated between each pair of parameters. c) Zoomed-in version of the density plot for parameters λ vs a

The most notable result from the above plots shown in Figure 11 is the high degree of correlation between estimations for λ (target cell replenishment rate) and a (infected cell death rate), with a correlation between the estimations of the two parameters of .88. This trend of highly correlated estimates for λ and a seen in the one subjects shown in Figure 11 was consistent among all individuals. It should be noted that the correlations as well as parameters distributions are all being calculated on the log base e scale. The implication of this high correlation observed between that λ and a is that the two parameters cannot be simultaneously estimated if we hope to have any certainty in their true values. This is because any pairings of the two parameters along their correlation line result in solutions that are equally favorable. This means that these values cannot be relied upon as parameter estimates, because of the dependency inherent in the parameter estimations of λ and a. As the goal of this thesis is to estimate parameter values in order to assess their biological meaning and significance, this issue of uncertainty as a result of correlation poses a problem. Thus a method had to be devised in order to overcome the inability to simultaneously estimate λ and a due to their high degree of correlation.

Repeating the fitting process with a fixed infected cell death rate

One method that could be implemented to resolve the inability of our model to simultaneously estimate both λ and a is first estimating the value a from other data sources, and then holding it constant in the final model fitting. Though up until this point the only data that

has been utilized is that from the rebound stage, which occurred after the cessation of treatment, we recall that the experimental data exists in two other stages: initial acute infection, and treatment. During the treatment stage, the subjects in the experimental group are all receiving ART at varying doses, in addition to the TLR7-agonist immunotherapy that select subjects are receiving (which isn't given until ART has already been administered for over a year). The way that ART functions is that it prevents HIV from infecting new uninfected cells. In terms of the differential equations governing the system, this means that the parameter β becomes zero. This means that the system of equations simplifies as follows:

$$\frac{dx}{dt} = \lambda - (d \cdot x)$$

$$\frac{dy}{dt} = -(a \cdot y)$$

$$\frac{dv}{dt} = (k \cdot y) - (u \cdot v)$$

In this system, the closed form solution for *x(t)*, *y(t)*, and *v(t)* can easily be calculated, where t_t is the time at which treatment began:

$$x(t) = \frac{\lambda}{d} - (\frac{\lambda}{d} - x(t_t))e^{-dt}$$

$$y(t) = y(t_t)e^{-at}$$

$$v(t) = \frac{v(t_t)(u \cdot e^{-at} - a \cdot e^{-ut}}{u - a}$$

Looking specifically at the solution for viral load, v(t), it is useful to note that $u \gg a$. This allows v(t) to simplify even further, to the following closed form solution during ART treatment:

$$v(t) = v(t_t)e^{-at}$$

This is useful to addressing the issue of inability to simultaneously estimate λ and a because it turns out that a, the death rate of infected cells, is relatively constant in an individual throughout the various stages of infection and treatment (Elemans et al 2011). This means that a can be estimated during the treatment stage by fitting the simple exponential equation above. The value of a obtained from this fit can then be fixed as a constant in the final fitting of the model to viral load data in the rebound stage. Fixing a thus addresses the problem of the correlation between λ and a because it allows for the value of λ to be estimated freely, independent of its correlation with a.

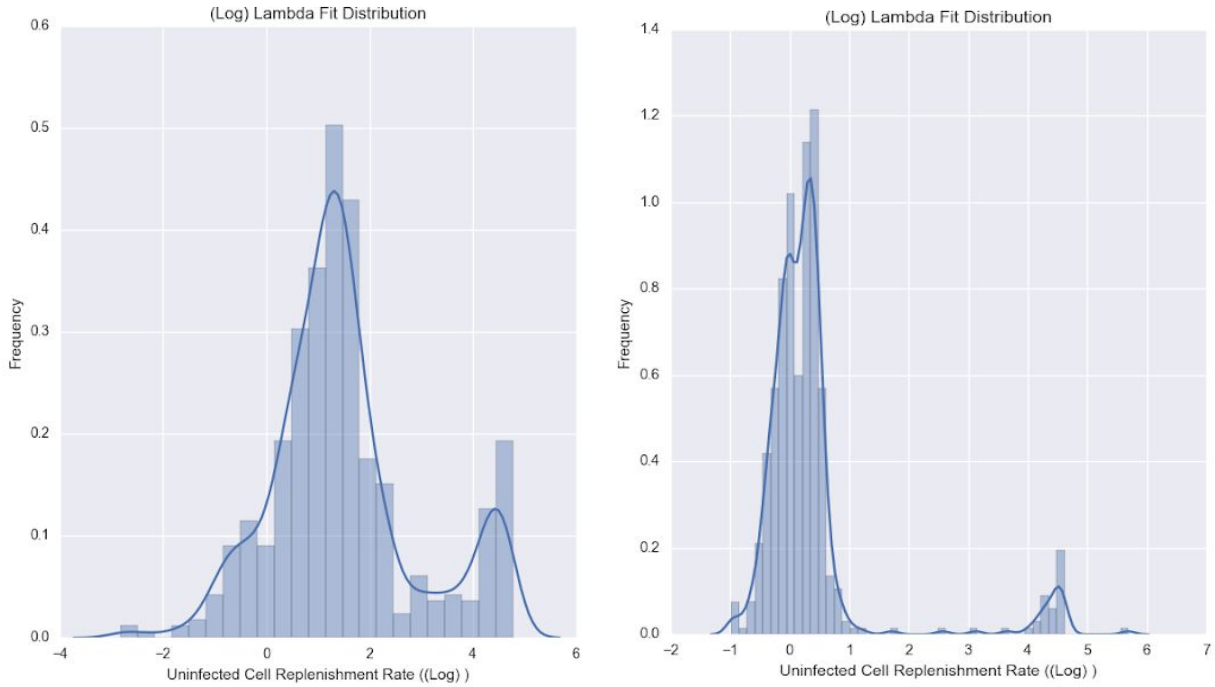


Figure 12: Sample distribution of λ before and after fixing *a*. subject 156-08 distribution of λ (trend seen across all individuals). a) histogram of λ values before a is fixed, b) histogram λ values with a fixed. The histograms reveal much more certainty in the estimation of λ after fixing a.

The distribution of the 500 best fit values for the various parameters that the fitting method converges upon, as shown in Figure 12 for subject 156-08, can serve as a proxy for parameter

uncertainty. If there is a high degree of variation in the distribution of best fit values for a particular parameter, that is an indication of higher levels of uncertainty in estimating the parameter value, as the implication is that many values of the parameter are comparably likely to arise from fitting the data. Here, for instance, in subject 156-08, the standard deviation of the distribution of log best fit λ values decreased from 1.43 when λ and *a* were simultaneously being estimated, to 1.06 when *a* was fixed and λ was estimated on its own. Similar trends of first fitting *a* on the treatment data increasing our certainty in λ estimations were observed for the vast majority of individuals, though it did not come without trade offs.

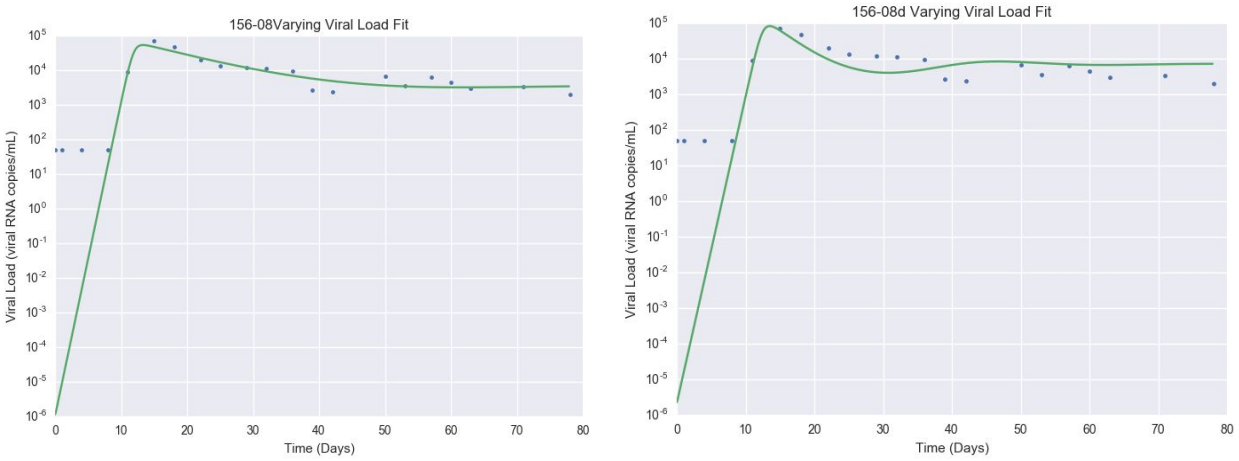
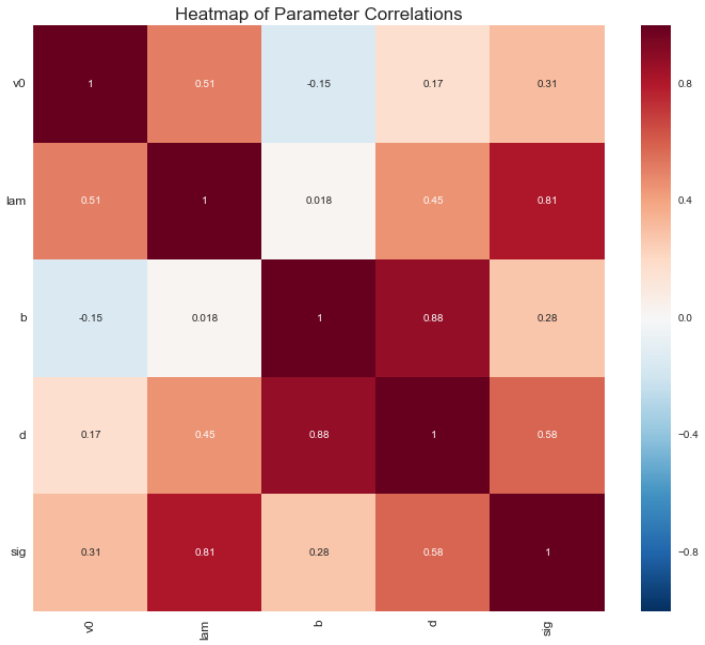


Figure 13: 156-08 comparison of rebound fit before and after fixing *a.* Observing visualizations of the fits with *a* varying (left) and fixed (right) the tradeoff between parameter certainty and fit accuracy is highlighted.

The primary tradeoff, as depicted in Figure 13, is between accuracy of the fit and thus our degree of certainty that these parameter values are truly representing the viral load data in these patients. To compliment the visualization, this can be quantified by calculating the AIC in both instances in order to ensure that fixing *a* wasn't drastically reducing the accuracy of the fit. AIC is used as a metric of model comparison because of the nature for likelihood, unlike sums of squared error,

to arbitrarily decrease as the number of parameters fitting the data increases. This is potentially problematic in cases such as the problem at hand, which involves comparison of models with different numbers of free parameters. AIC accounts for this by measuring the accuracy of the fit, while at the same time adjusting for the number of parameters being estimated.

There were different results across individuals of fixing the death rate of infected cells: in some cases in reduced the AIC, in others in raised it. In the case of the example individual shown above, the AIC was 20.05 before fixing *a,* and 29.57 with *a* fixed. Recall that a lower value for AIC is preferred, and thus the model in which *a* is allowed to vary results in a better fit for the data. This is evidenced by looking at the fit on the data visually, as seen in Figure 13. In most cases this change was slight, indicating that fixing the α parameter didn't have a huge bearing in the ultimate success of the model to fit the data. Table 2 for reports completely on how fixing *a* by first fitting on the treatment data impacted AIC, as well as parameter uncertainty.



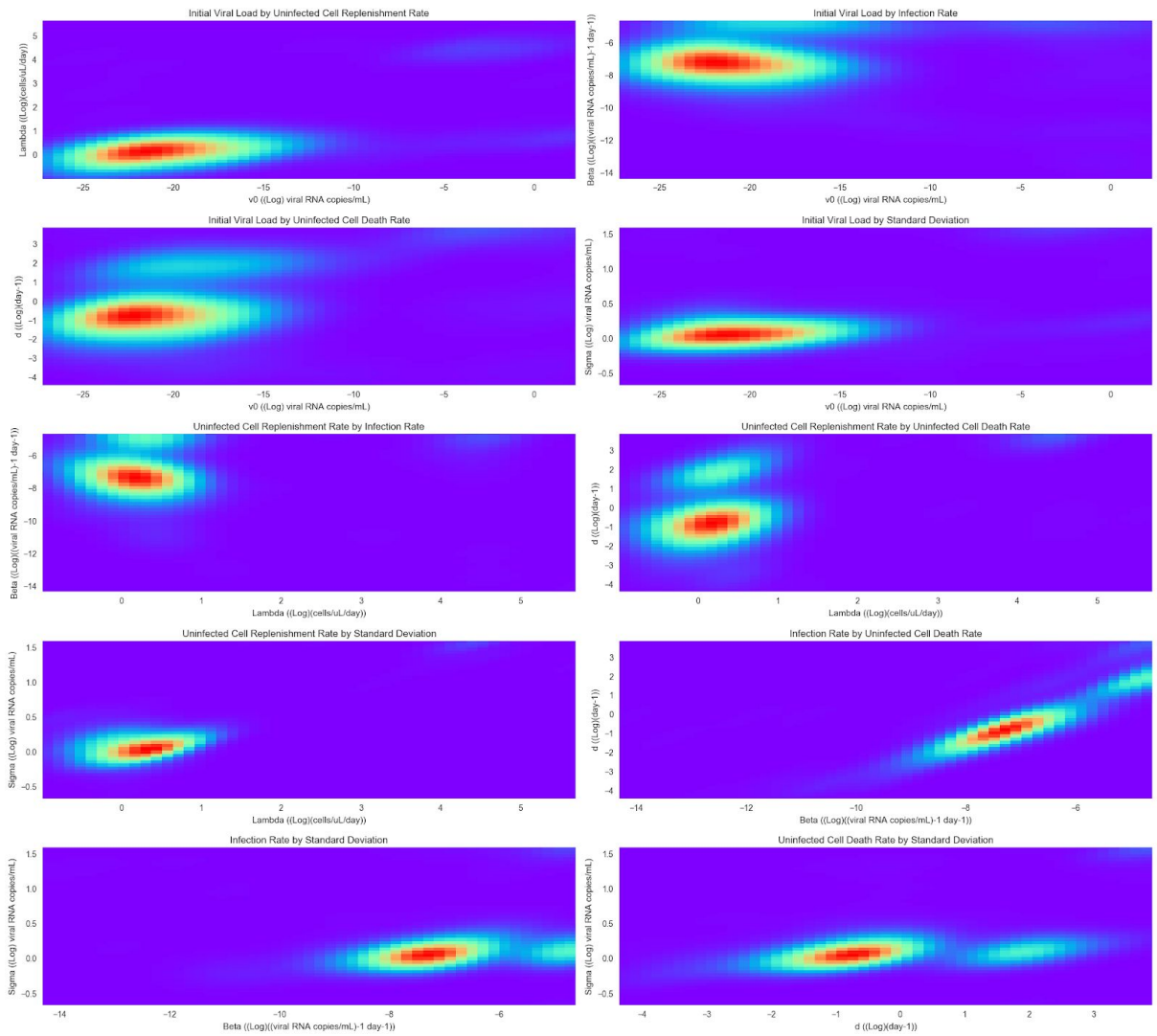


Figure 14: 156-08 Correlation plots when *a* is fixed from treatment data. a) The Pearson correlation coefficient calculated between each pair of parameters. Darker color and a value closer to 1 indicates parameters that are more correlated. b) On each kernel density estimation plot the x-axis is one parameter and the y-axis is another, and the color represents the frequency of a particular combination of these parameters in an optimal parameter set returned by lmfit.

Addressing non identifiability by fixing uninfected cell death rate

Looking at the correlation results from the fits in which v_0 , λ, β, and d were simultaneously estimated, and a was fixed based on first fitting it to the treatment data, it becomes clear that while fixing a corrected for one parameter dependency, there are still issues

of dependency, and thus of ability to simultaneously identify, β and d. An example of these dependencies for subject 156-08 is depicted in the correlation plot and heatmap shown in Figure 14. This trend was relatively constant across individuals. Remarkably, the correlation coefficient between β and d under these circumstances is .88, meaning that β and d are equivalently difficult to simultaneously estimate as λ and a were in the initial fit of the rebound data. In order to be able to be able to maximize certainty in our parameter estimates, β and d therefore cannot be simultaneously estimated. I attempted to address this issue by fixing d to .05 (day^{-1}), a value that has been has been experimentally for memory CD4+ T cells in macaques and was found to be both relatively constant across individuals (de Boer et al, 2003). Thus I again implemented the same maximum likelihood fitting algorithm on the rebound data from phases 1 and 2, this time with a fixed at the value obtained from fitting to the treatment data, and with d held constant at .05 for all individuals. The visualizations for two sample solution plots are depicted in Figure 15. This provides indication qualitatively that addressing the final issue of parameter identifiability by holding d constant at .05 doesn't necessarily hinder the fit accuracy substantially, and does increase parameter certainty, as desired. Because all population-wide issues of identifiability at this point have been resolved, *and* preliminary visual inspection seems to indicate that the fits are still fitting the underlying patterns in the viral load data, this presents itself as a plausible method for fitting the rebound viral load data.

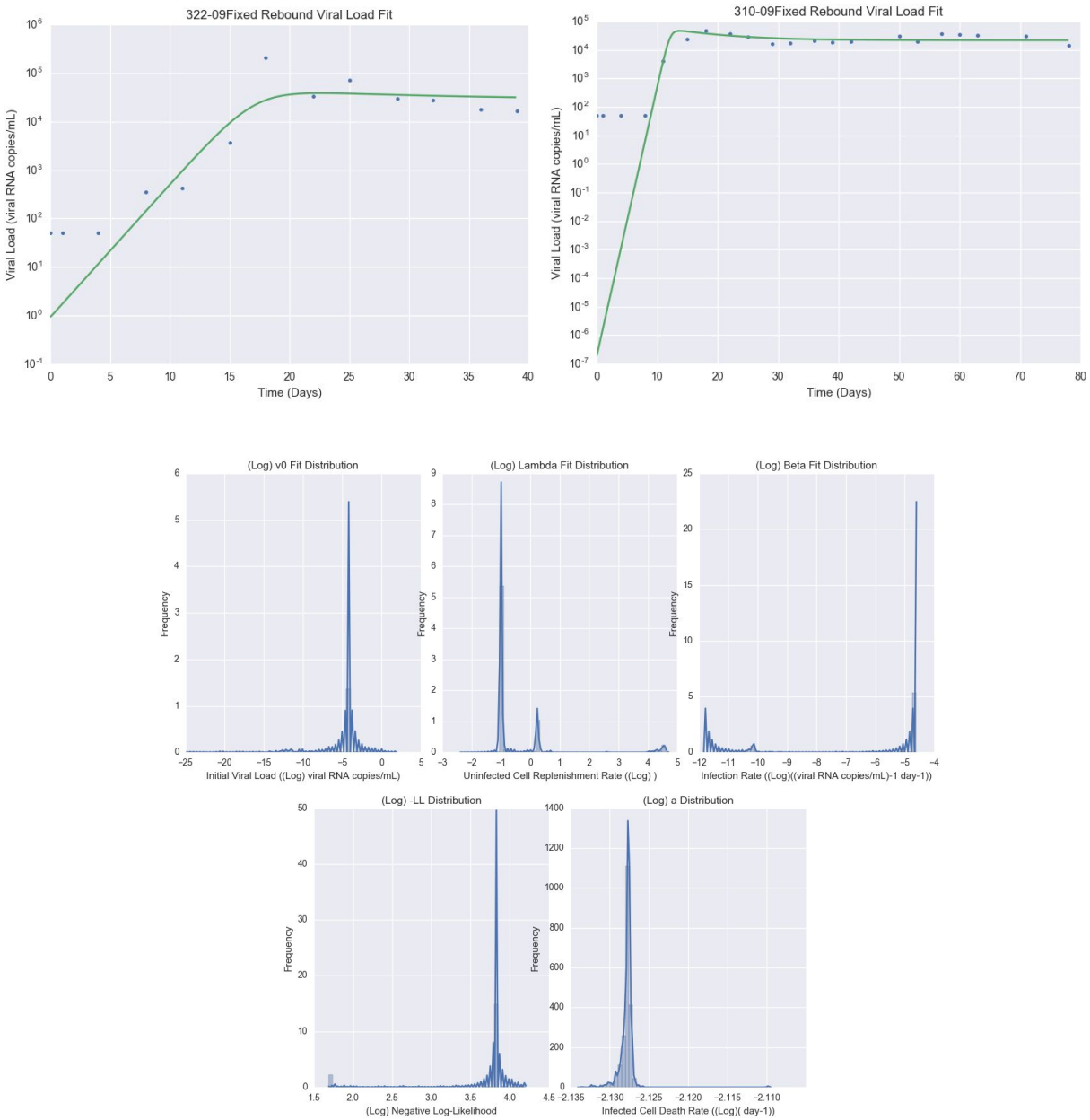


Figure 15: a) Sample plots of best fit parameter solution when *a* and *d* are both fixed. 322-09 and 310-09. See Appendix A for plots of all fits. **b.) Histogram of parameters when *a* and *d* are fixed.** Highlights high degree of certainty of parameters being estimated (v_0 , λ , β, negative log likelihood, and *a* fit distribution from treatment data)

Assessing trade offs: final rebound fits

The results of all three cases of fitting (with only *u* and *k* fixed, with *u, k,* and *a* fixed, and with *u, k, a,* and *d* fixed) on the rebound data are summarized in the table below. For each patient

it compares the three different variations of the maximum likelihood model fitting method across two dimensions: fit accuracy and parameter uncertainty. Fit accuracy was represented simply by AIC. Average uncertainty is a bit more complicated, as it weights the standard deviation of the distribution of fit results for all parameters being estimated. The purpose of this is to quantify the overall uncertainty combined in estimating *all* parameters, rather than making individual comparisons between uncertainty in particular parameter estimation. This was done by first calculating the ratio of the standard deviation divided by the best fit value for each parameter (to weight uncertainty proportional to scale), where the standard deviation is of the data on log base *e* scale, and the best fit value for the parameter is also logged. Because these quantities are on the log scale and thus many were initially negative, these quantities are first squared, and then summed up. The square root of this sum is taken to normalize, and then we divide by the number of parameters being estimated, to ensure that there is not more uncertainty simply by virtue of estimating more parameters, but that the individual parameter estimates themselves are what is dictating the uncertainty estimate. Formally, for each parameter being estimated θ_i for $i \in n$ where $\hat{\theta}_i$ is the best fit value for θ_i :

$$\frac{\sqrt{(\frac{STD(\theta_1)}{\hat{\theta_1}})^2 + (\frac{STD(\theta_2)}{\hat{\theta_2}})^2 + ... + (\frac{STD(\theta_n)}{\hat{\theta_n}})^2}}{n}$$

To summarize these major findings of the model fitting to the rebound data and ultimately decide which set of parameters best describe the data, we thus look at these metrics in Table 2 to weight the likely trade off between parameter uncertainty and fit accuracy. The blue highlighted boxes indicate the version of the fit that produces the "best" fit for each individual,

as measured by the lowest AIC and the lowest uncertainty value. Some of the calculations for uncertainty result in NaNs being produced by virtue of the fact that logs were taken on values potentially close to zero, and the added potential for any $\widehat{\theta}_i$ (as it typically is for β or v_0 for example).

Table 2: Summary of Rebound Fits

	Accuracy of Fit (AIC)			Average Uncertainty		
Individual	None fixed	*a* fixed	*a* and *d* fixed	None fixed	*a* fixed	*a* and *d* fixed
105-09	20.74	23.26	22.04	NA	0.249	0.19
156-08	20.40	29.58	28.42	0.229	0.677	0.98
162-09	35.3	27.76	34.95	0.211	0.928	1.01
166-08	24.8	24.93	23.39	0.316	0.302	0.36
205-08	27.99	31.4	30.86	0.245	0.717	0.629
234-09	14.87	17.66	20.76	NA	0.452	0.124
267-08	23.39	23.33	21.34	3.736	40.823	11.862
280-09	23.1	21.76	20.23	0.201	2.265	1.814
280-10	32.77	31.66	29.6	0.286	0.119	0.103
288-10	32.98	31.82	33.98	0.337	0.233	0.206
293-09	25.77	22.64	20.31	NA	0.109	0.095
295-10	28.76	31.55	33.43	0.331	2.048	2.345
304-10	25.2	23.41	31.96	0.476	0.356	0.185
305-10	28.01	26.86	25.05	0.476	0.418	0.190
310-09	16.35	14.39	13.42	0.573	0.640	1.469
322-09	23.42	21.74	20.93	111.666	25.480	27.545
341-10	23.17	21.75	20.10	0.745	nan	1.663

374-09	8.2	8.89	10.21	1.956	2.850	1.116
412-10	24.3	20.95	24.71	0.530	0.183	0.161

The quantities in the table overall match what can be seen with the eye when looking at these fits: that there is not one single identifiable pattern as to which fit of the three methods will be the "best". It is not as simple as recognizing that fixing *a* will always have a particular impact on the accuracy of the fit, fixing both *d* and *a* will have some separate impact, etc. The impact of these various forms of model fitting instead varies from individual to individual. On aggregate, however, when looking at the the fitting algorithm that minimizes AIC and uncertainty in parameter estimation, the method with the fewest parameters being simultaneously estimated appears to be the optimal solution for both accurately and precisely fitting rebound viral load data.

The pattern of AIC results is somewhat unexpected, because we see that the model with both *a* and *d* fixed most frequently had the lowest AIC. This is surprising because there were no instances in which fixing *a* and *d* improved overall fit accuracy, it either stayed roughly the same or got slightly worse. This is highlighted by Figure 16, which illustrates the comparison in fit between the solution where *a* and *d* both vary, and the solution where they are both fixed for one of the individuals that had the lowest AIC for the fit in which *a* and *d* were both fixed. We are likely seeing this result by virtue of the fact that AIC penalizes for more parameters being estimated inherently. Thus while the model with more parameters being estimated may resemble the data more closely, it is receiving a penalization factor for each additional parameter being estimated. It seems to be a reasonable conclusion then that the fits where there are more free

parameters would have to be substantially better fits of the data in order to make up for their inherent disadvantage by way of AIC model comparison by estimating more parameters.

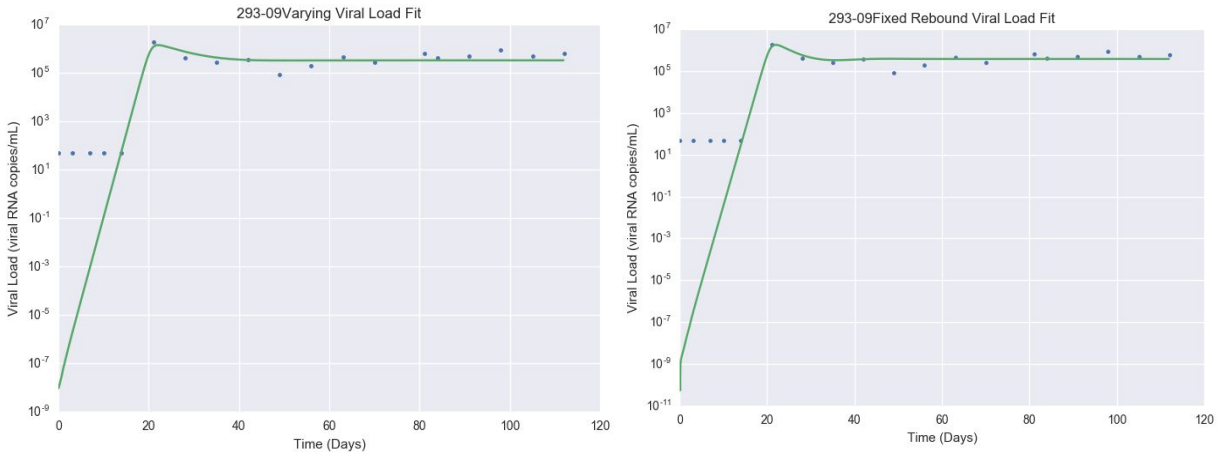


Figure 16: Plot of 293-09 before and after fixing *a* and *d.* This is a visualization of one instance in which estimating fewer parameters resulting in a lower AIC.

Visualizations of the plots as well as the AIC values make it clear that the fits are relatively comparable in the case where *a* and *d* are fixed as to when all are allowed to vary for many individuals. Though this is certainly not the case for all individuals, it is true that fixing *a* and *d* never makes the fits substantially worse. In some of the instances in which the fitting method with *a* and *d* fixed is clearly missing the peak viral load or converging on a suboptimal solution for some other reason, this trend was also observed in the results of the parameter estimates in which *a* and *d* were also being fit. This can be seen qualitatively (Appendix A) by the visualizations of these fits, and is quantified by the AIC values that are more often than not better when *a* and *d* are fixed. Additionally, the distributions of the parameters being fit reveal that there is in fact more uncertainty when more parameters are being simultaneously estimated.

Therefore fixing *a* and *d* produces fits to the data that are comparable and even often better than those produced by fitting all five parameters simultaneously and increases our certainty in these estimates. Because of this, the parameter values that resulted from holding *a* and *d* constant are the preferred solution, and the most reliable parameter estimates as explored thus far. Table 3 contains the best fit parameter values using this fitting method.

Table 3: Final Rebound Parameter Estimates

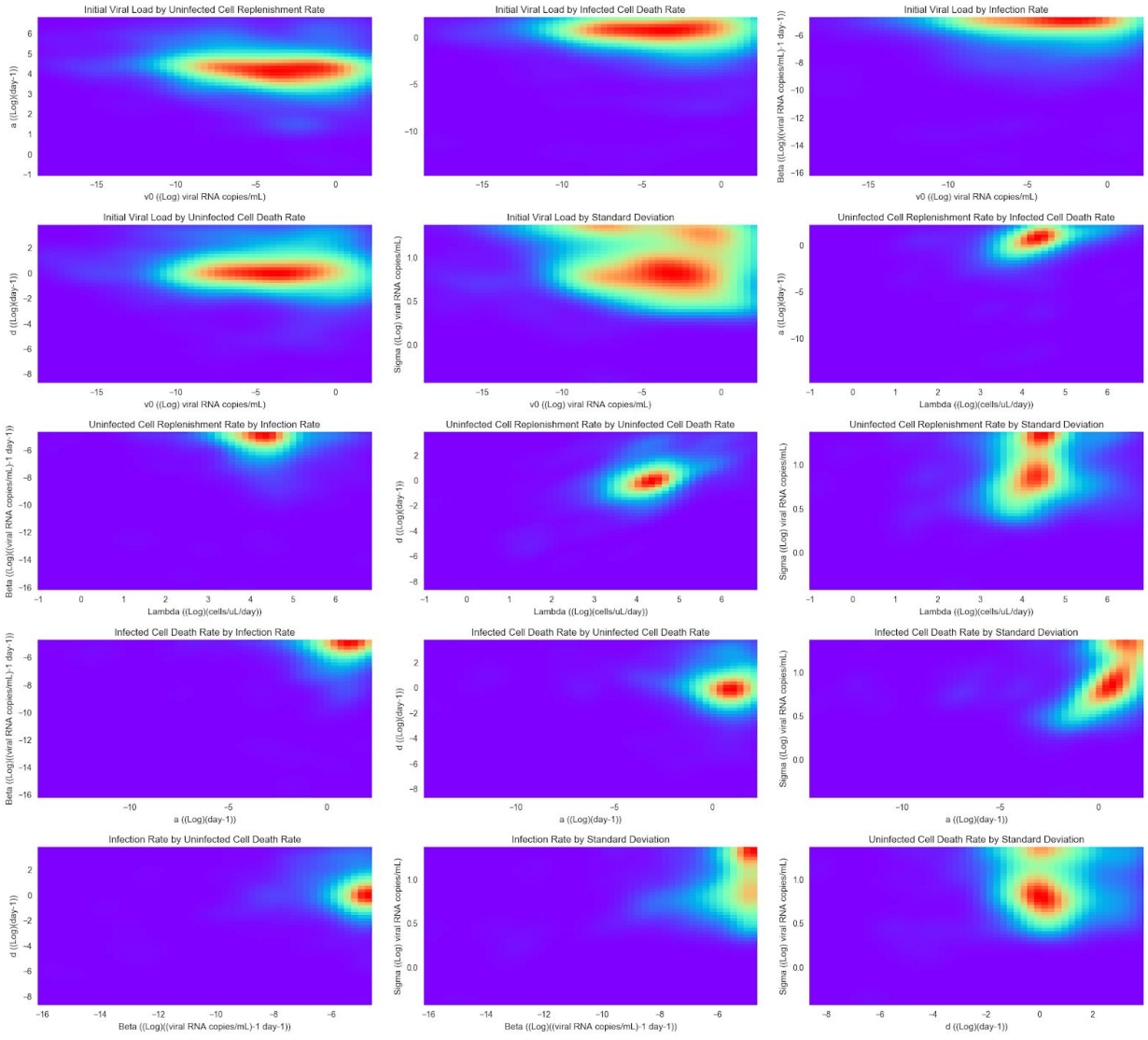
Subject ID	v_0	λ	a	β	σ
304-10	9.65E-06	3.32	0.127	1.56E-05	2.92
288-10	4.44E-05	3.817	0.217	1.26E-05	3.37
295-10	0.0743	1.12	0.285	2.41E-05	3.16
162-09	5.47E-07	0.757	0.203	5.94E-05	4.22
341-10	2.21E-07	10.39	0.686	5.62E-06	1.95
305-10	0.034	17.68	0.482	1.99E-06	1.98
412-10	0.0022	19.19	0.229	1.83E-06	2.38
293-09	1.05E-09	71.95	0.350	6.95E-07	1.77
280-10	1.13E-07	12.39	0.412	6.04E-06	2.54
280-09	6.45E-08	0.819	0.101	7.35E-05	1.64
374-09	0.505	12.47	0.204	1.94E-06	1.35
166-08	0.0039	2.516	0.361	1.49E-05	1.78
156-08	7.26E-07	1.327	0.364	4.64E-05	2.23
205-08	9.99	8.949	0.491	1.74E-06	2.424
267-08	0.945	19.817	0.315	1.16E-06	1.66
234-09	1.15E-05	63.20	0.112	6.84E-07	2.67
105-09	0.0036	11.95	0.077	2.46E-06	1.72
310-09	7.14E-07	1.26	0.119	4.32E-05	1.40
322-09	0.947	1.128	0.077	1.48E-05	2.39

Fitting data from acute infection

Since one of the unique features of the data from this study was the availability of information on acute infection, it could also be informative to estimate parameter values for the acute stage. To do so, we can apply the same principles used to fit the rebound data, and the same process of selection. The data follows the same format, and is censored below 50 viral RNA copies/mL, and thus will benefit from the same implementation of the maximum likelihood approach to fitting the viral load data that was used on the rebound data. Similar to what I did with the rebound data, the first step in attempting to estimate the parameter values on the acute data is fixing only u and k at their known values, and allowing all other parameters to be fit simultaneously. Here I looked for similar issues of correlation between parameters being estimated that would again be indicative of inability to simultaneously estimate a certain subset of the parameters. Due to drastically different kinetics in the acute stage of infection, there is no reason to necessarily assume that the same estimation dependencies will exist, and thus to assume that a and d will need to be fixed in order to insure a high degree of certainty in the parameter estimates.

Though there certainly were non-negligible correlations seen between best value parameter estimates within individuals, there were no pervasive trends of strong correlation between the same parameters in every, or even most, individuals. Rather than highlighting issues of simultaneous estimation, the results of the correlation heatmaps for the acute fitting indicated a separate issue that is depicted in Figure 17: a high degree of uncertainty in estimating v_0. There is interestingly a high degree of certainty in measuring the other parameters, as is evidence by

the very small regions of high density of solutions shown in Figure 17. Because the problem in fitting the acute data is thus not one of inability to identify as a result of correlation, the solution used to improve the fit on the rebound data will not be applicable.



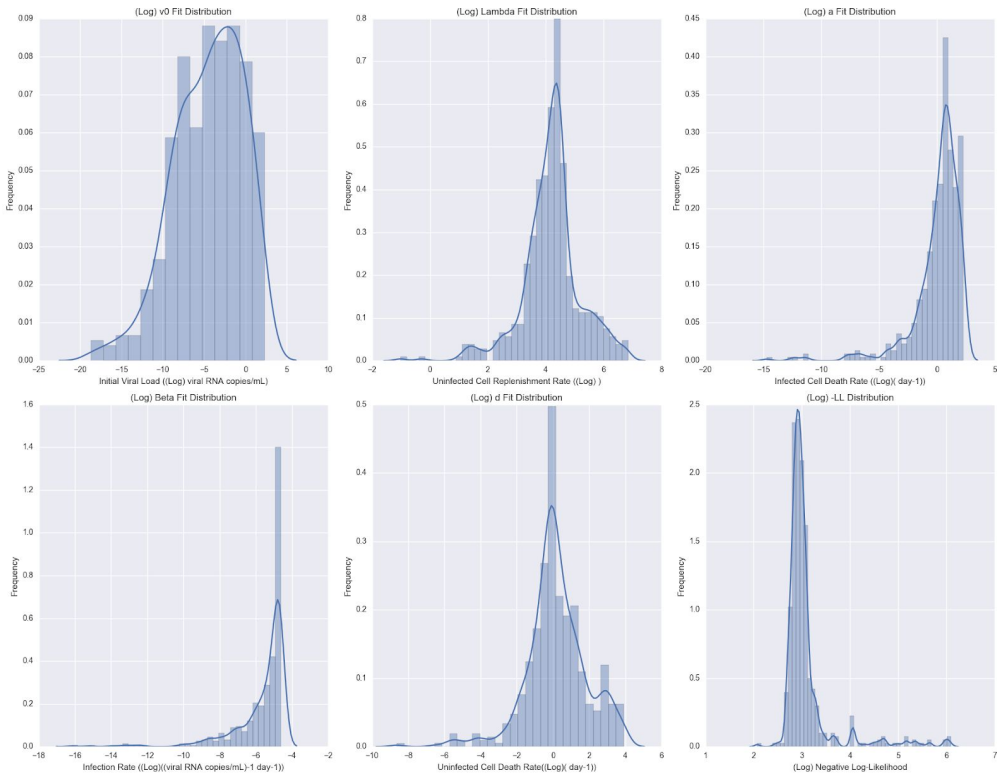


Figure 17 : 267-08 Acute fit kernel density estimate correlation plot and parameter estimation uncertainty. Both visualizations of the data highlight the high degree of uncertainty in estimating v_0. The KDE plot shows that all parameters being estimated are not burdened by issues of dependence, and thus should be simultaneously identifiable

The most important result here is the high degree of uncertainty in estimating v_0, as is evidenced by its large region of high density values, meaning that there is a very large range of data points that are equally likely to be the best fit value for v_0. This is evidenced by the pictured distribution of v_0 on the log scale, and the standard deviation of this distribution of 4.16 on the log scale. The most likely explanation for this points back to one of the early limitations of the fitting model identified during sensitivity testing: the difficulty of fitting the model to viral load data when there are few time points on the upslope of the data. The acute data was measured more infrequently than the rebound, with measurements only being taken every seven days as opposed to every three days in the rebound data. For many of these patients, that meant having

only one measurement on the upslope of the viral load trajectory before reaching the peak. This particular weakness of our model to accurately fit the upslope of the data when there are few early time points is likely what is causing the estimates for v_0 to be so uncertain. Figure 18 highlights this relationship, showing two examples of fits in which the upslope is reasonably approximated, and two examples of instances in which it was not. For future studies, the implication of this is that frequent viral load measurements, particularly early enough in the course of the infection to obtain a more granular picture of the viral upslope leading to the peak, is crucial is the data is to be used for mathematical modeling.

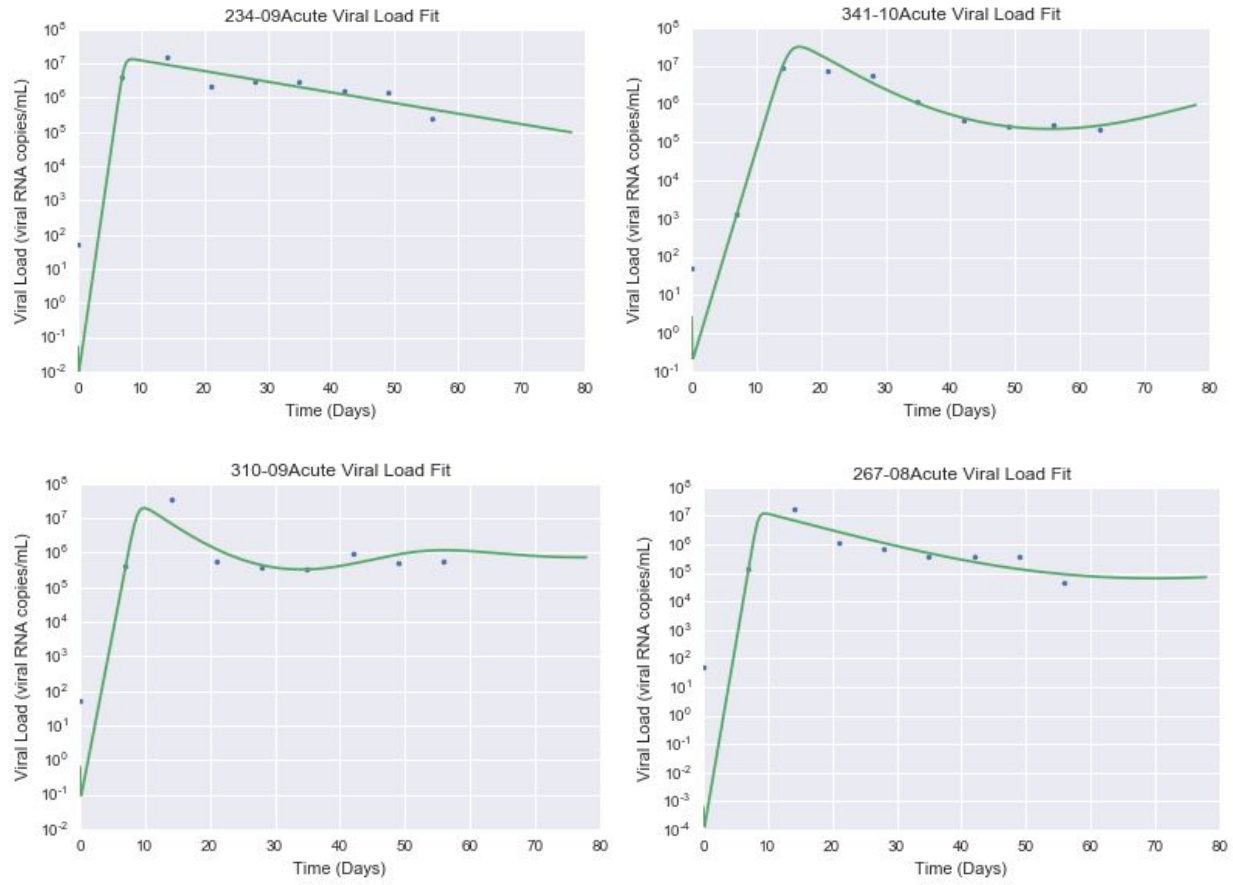


Figure 18: Sample plots of fits on the acute data. a) and d) serve as examples for the subset of the acute fits that have a very steep early exponential growth rate, a sharp peak, and then a relatively linear fit to the rest of the data. b) and c) serve as examples of the subset of the acute fits that have high amplitude oscillations in equilibrium, and converge on an unrealistically high value for v_0.

Table 3: Final Acute Parameter Values

Subject ID	v_0	λ	a	β	d	σ
105-09	0.5770	884.58	0.0123	0.00029	0.3258	2.667
374-09	9.3817	9.8662	0.1371	1.13E-07	0.0012	1.585
322-09	0.0040	49.45	0.2656	8.93E-08	0.0028	1.323
156-08	0.0005	330.425	1.3209	0.010000	0.4621	1.778
310-09	0.6332	135.959	0.3106	9.17E-08	0.0097	2.124
234-09	0.0536	0.2727	0.0719	2.30E-07	3.99E-05	1.673
205-08	1.3E-3	43.6	0.2760	1.55E-07	4.9E-3	2.18
267-08	6E-4	10.3	0.1381	2.51E-07	1.6E-3	1.92
166-08	3.2E-3	87.9	0.3661	9.40E-08	6.0E-3	1.54
280-09	10.00	248	0.3504	2.77E-08	9.8E-3	1.43
162-09	1.49E-06	9.3	0.089	2.88E-07	1E-3	1.72
280-10	9.99	508.13	0.2821	1.22E-08	7.3E-3	1.55
295-10	0.062	1.49	0.1433	4.64E-08	6.11E-05	2.95
177-10	2E-04	49.8	0.1870	5.80E-07	0.015	6.167
288-10	0.72	46.3	0.0469	0.007322	2.11	3.58
412-10	0.69	889	0.4369	0.006164	21.5	3.46
304-10	2.37	597	0.1785	1.09E-08	9.5E-3	1.44
344-10	2.87	150	0.2746	3.00E-08	3.8E-3	2.08
293-09	10.0	64.7	0.2296	6.02E-08	6E-3	1.58
341-10	2.64	161	0.2423	2.95E-08	6.4E-3	1.36
305-10	3.03E-05	3.6550	0.0529	2.56E-07	1E-3	3.77

Overall fitting on the acute infection data was not nearly as successful as was fitting on the rebound infection data. Figure 18 a) and d) show the convergence on unrealistically steep early exponential growth rate seen in some individuals, followed by a very jagged decrease. Though some of the fits at first glance may appear to be fitting the data well, for example Figure 18 b) and c), closer examination shows that even when the fit closely resembles the data, there are some nonsensical trends that pervade. These include a very high v_0 followed by an initial dip down, which is not how we would expect viral load to behave upon infection. Additionally, the

equilibrium that the solution converges on is meant to reflect dampened oscillations. Here, however, we are seeing oscillations with an incredibly high amplitude, that while they may fit the viral load data locally it is unlikely that this oscillatory nature of viral load as $t \rightarrow \infty$ will be an accurate representation of the data. Because of these systematic issues in the fitting of the model in its current form to the acute infection data, including the two above mentioned trends and the inability of the fitting method to estimate initial viral load, it is unlikely that we can trust the best fit parameters to properly represent the viral load data. Potential explanations and steps to be taken to improve the fit on acute infection data are discussed further in the discussion. This is an unfortunate result given the fact that acquiring HIV acute infection data is so rare in and of itself, and that there were many questions that I hoped to answer in terms of biological implications of the estimated parameter values.

Final analysis and biologically significant implications

Despite shortcomings in the fitting results, particularly on the acute data, it is still useful to explore the ties back to the biological implications of HIV infection, treatment, and potential cure to see if any substantial trends exist in the data. To do this I explore correlations between acute and rebound parameter estimates, associations between acute parameter estimates and whether or not a patient is ultimately cured (as 177-10 and 344-10 ultimately were), implications of composite parameters that can be calculated from the fit of the model to the data, as well as differences in parameter estimates across treatment groups (i.e. those that received the TLR7-agonist immunotherapy and those that did not).

In terms of looking at pairwise correlations between parameter values in during acute infection versus those in rebound, there were no statistically significant correlations between the same parameter value during the two stages of infection, or with between any of the parameters and v_0. This was an unfortunate result, but not surprising given the lack of certainty in the parameter estimates for the acute data, as well as the incredibly small sample size over all (just $n = 19$ when the two treated individuals who were cured are excluded). For the subjects who were cured, unfortunately the consequence of their sample size being so small ($n = 2$) there was no way to statistically evaluate whether or not there was a difference between parameter values in acute infection that ultimately give rise to a successful treatment, and those that don't. This meant that the only option was visually seeing if anything about these individuals was different in the acute infection in terms of their parameter values. For both 177-10 and 344-10, none of their parameter values were outliers. Figure 19 additionally highlights that there was nothing particularly noteworthy about their viral load trajectory in acute infection.

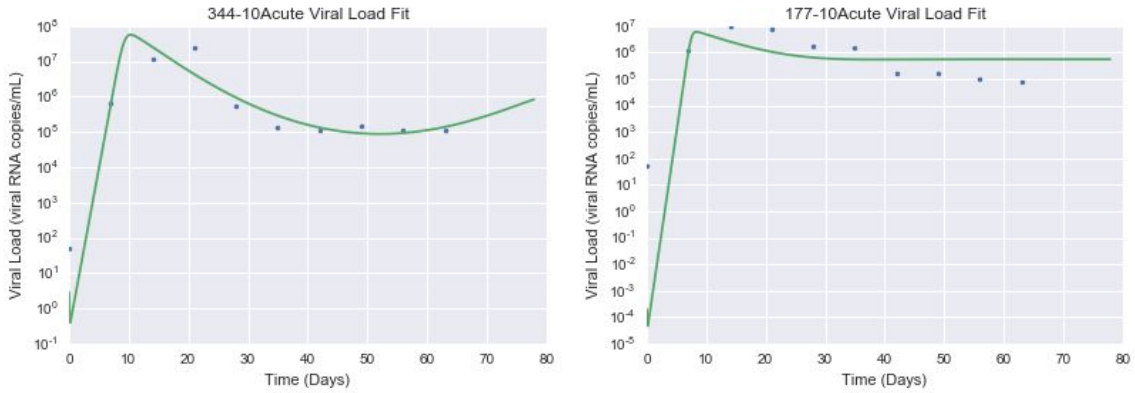


Figure 19: Plots of acute infection and fit for two treated individuals. There is nothing upon preliminary inspection that separates the acute infection for these two individuals from the rest of the individuals.

Another bit of analysis that didn't return significant results but is still worth mentioning is that when doing a two sample t test to compare group means in the rebound infection stage

between the group that received the TLR7-agonist in addition to ART with the group that received a placebo in addition to ART between the distribution of each parameter value, there were no statistically significant differences between any of the parameters (i.e. *a* in the placebo group compared with *a* in the treatment group, and so on for all parameters being estimated).

The final bit of analysis in looking at biological relevance of the results of this process was in computing several relevant composite parameters for both rebound and acute infection. The values calculated include value and time of maximum predicted viral load, value and time of maximum observed viral load, R_0, r_0, the set point equilibrium viral load, difference between peak viral load and set point, as well as the target population initial condition estimate ($\frac{\hat{\lambda}}{\hat{d}}$). For the acute infection data I also calculated the area under the curve, as this is indicative of the total amount of virus accumulated throughout the infection. These results are listed in Tables 3 and 4, and some of the implications of these composite parameter values are discussed in the discussion.

Table 4: Rebound infection composite parameter estimates

Subject ID	Max predicted	t(max predicted)	Max observed	t(max observed)	R_0	r_0	v^*	Initial target population	Max - SPVL
280-10	3.37E+05	12.4	2.63E+05	14	7.89	2.84	5.71E+04	247.9	2.06E+05
305-10	2.62E+05	18.2	2.94E+05	21	3.17	1.05	5.47E+04	353.5	2.39E+05
280-10	3.37E+05	12.4	2.63E+05	14	7.89	2.84	5.71E+04	247.9	2.06E+05
288-10	1.16E+05	14.2	1.81E+06	21	9.67	1.88	3.43E+04	76.3	1.78E+06
293-09	1.78E+06	22.0	1.90E+06	21	6.22	1.83	3.75E+05	1438.9	1.52E+06
295-10	2.20E+04	17.6	2.99E+05	21	4.12	0.89	6.47E+03	22.4	2.93E+05
304-10	1.20E+05	13.7	1.19E+06	21	17.73	2.13	5.35E+04	66.5	1.14E+06
341-10	1.73E+05	17.5	8.67E+04	21	3.70	1.85	2.40E+04	207.8	6.27E+04
412-10	5.07E+05	17.9	5.31E+06	21	6.68	1.30	1.55E+05	383.8	5.16E+06
162-09	2.30E+04	16.8	5.05E+05	21	9.62	1.75	7.26E+03	15.1	4.98E+05
305-10	2.62E+05	18.2	2.94E+05	21	3.17	1.05	5.47E+04	353.5	2.39E+05

280-10	3.37E+05	12.4	2.63E+05	14	7.89	2.84	5.71E+04	247.9	2.06E+05
288-10	1.16E+05	14.2	1.81E+06	21	9.67	1.88	3.43E+04	76.3	1.78E+06
293-09	1.78E+06	22.0	1.90E+06	21	6.22	1.83	3.75E+05	1438.9	1.52E+06
295-10	2.20E+04	17.6	2.99E+05	21	4.12	0.89	6.47E+03	22.4	2.93E+05
304-10	1.20E+05	13.7	1.19E+06	21	17.73	2.13	5.35E+04	66.5	1.14E+06
341-10	1.73E+05	17.5	8.67E+04	21	3.70	1.85	2.40E+04	207.8	6.27E+04
412-10	5.07E+05	17.9	5.31E+06	21	6.68	1.30	1.55E+05	383.8	5.16E+06

Table 5: Acute infection composite parameter estimates

Subject ID	Max predicted	t(max pre)	Max observed	t(max obs)	R_0	r_0	v^*	Initial target population	Max - SPVL	AUC
105-09	8.1E+07	55.9	1.7E+07	14	1.4E+05	1711.7	1.6E+08	2.7E+03	-1.4E+08	2.7E+09
156-08	2.3E+77	21.3	2.4E+07	14	1.2E+04	15543.5	5.4E+05	7.2E+02	2.4E+07	2.3E+76
162-09	1.7E+07	8.1	1.5E+07	14	60.2	5.2	2.2E+05	8.5E+03	1.5E+07	2.2E+08
166-08	2E+07	11.6	1.2E+07	14	8.2	2.6	4.6E+05	1.5E+04	1.2E+07	9.3E+07
177-10	6.0E+06	8.3	9.9E+06	14	22.8	4.1	5.5E+05	3.4E+03	9.4E+06	6.6E+07
205-08	1.3E+07	11.5	3.3E+07	14	10.8	2.7	3.1E+05	8.9E+03	3.3E+07	7.6E+07
234-09	1.3E+07	8.5	1.5E+07	14	47.5	3.3	8.1E+03	6.8E+03	1.5E+07	2.0E+08
267-08	1.2E+07	9.5	1.7E+07	14	26.2	3.5	1.6E+05	6.6E+03	1.7E+07	1.1E+08
280-09	2.4E+07	17.8	2E+07	21	4.4	1.2	1.2E+06	2.5E+04	1.9E+07	1.7E+08
280-10	8.4E+07	14.5	7.2E+07	14	6.5	1.6	3.3E+06	6.9E+04	6.9E+07	5.9E+08
288-10	2.0E+06	62.9	2.1E+07	14	7440.7	349.2	2.2E+06	2.2E+01	1.9E+07	9.2E+07
293-09	1.3E+07	17.3	2E+07	21	6.2	1.2	5.1E+05	1.1E+04	1.9E+07	1.11E+08
295-10	4.1E+07	12.2	3E+08	14	17.2	2.3	2.1E+04	2.4E+04	3E+08	3.7E+08
304-10	8.6E+07	18.6	5.7E+07	21	8.3	1.3	6.4E+06	6.3E+04	5E+07	9.4E+08
305-10	7.2E+06	17.3	5E+07	28	39.4	2.0	1.5E+05	3.7E+03	5E+07	1.5E+08
310-09	2E+07	9.9	3.5E+07	14	9.0	2.5	8.5E+05	1.4E+04	3.5E+07	1.3E+08
322-09	2.7E+07	10	1.2E+07	14	12.8	3.1	3.7E+05	1.8E+04	1.2E+07	1.5E+08
341-10	3.11E+07	16.6	8.8E+06	14	6.7	1.4	1.2E+06	2.5E+04	7.6E+06	2.5E+08
344-10	5.7E+07	10.3	2.4E+07	21	9.5	2.3	1.1E+06	4E+04	2.3E+07	3.3E+08
374-09	1.4E+07	11.2	7.8E+06	21	14.9	1.9	1.5E+05	8.3E+03	7.7E+06	1.4E+08
412-10	4.4E+06	52.3	7.3E+07	14	1271.7	555.1	4.4E+06	4.2E+01	6.9E+07	2.7E+08

Discussion

In this thesis I have developed and applied methods to fit a system of differential equations describing the kinetics of HIV/SIV infection in the body to data from acute infection, treatment, and eventual viral rebound when treatment is stopped. This work has contributed important conclusions about our ability to estimate the parameters parameters from data routinely collected in pre-clinical studies of new HIV cure interventions. In particular, I examined the optimal number of parameters that can be simultaneously estimated from the data, prioritizing both the overall accuracy of the model fit, and the level of certainty in the parameter estimates. I found that only after fixing three of the seven possible model parameters (viral burst rate k, viral clearance rate u, infected cell death rate a) using values from separate experimental studies, and separately fitting uninfected cell death rate d using data during ART, could I simultaneously identified all the remaining parameters during the rebound phase. This is an important result for the future of mathematical modeling of HIV and attempts to recover estimates for the true unknown parameters. This places limits on our ability to understand the mechanism of action of any drug meant to alter viral rebound. The benefit of using mechanistic mathematical models (as opposed to regular statistics) is that the parameter values obtained from the data have clear biological implications. In addition to the basic model parameters, we can also calculate other composite parameters with intuitive biological meanings, as listed in Tables 4 (rebound) and 5 (acute). For example, the average lifespan of infected cells is given by $1/a$. I obtained two different a estimates: one coming from the acute infection data directly, and one from the treatment data (so as to eventually be held as constant in the rebound data fitting). The

median estimated value for this quantity across all subjects was almost identical in acute vs in rebound: 4.37 days in rebound and 4.35 days in acute infection. This makes sense given the assumption that was made in order to fix *a* with its value from the treatment data in the first place: that this parameter should be relatively constant regardless of stage. Note that we look at the median rather than the mean because of outliers that skew the summary statistic. The implication of this is that there is little difference in the life expectancy of a cell once infected with HIV, regardless of treatment. We can also use our parameter estimates to look at concentration of target cells, and what this says about whether or not all CD4+ cells are potential targets of HIV. The initial population of target cells can be estimated by the ratio of λ/d in the acute infection. On aggregate, the average of this ratio across all of the acute infection data is $1.6*10^4$ cells/mL and for rebound just 253 cells/mL. This is important because both values, especially that of rebound, are much lower than the estimated total CD4+ T cell population that has been experimentally measured to consistently be on the order of 10^6 cells/mL. The implications of this are that *all* CD4+ T cells may not be the target population for HIV as has traditionally been assumed, but rather that some subset of them are. Particularly, this suggests that an even smaller subset of CD4+T cells might be target cells of HIV upon cessation of ART. Future research could be directed towards understanding this further, and understanding what distinguishes CD4+ T cells that are and are not potential targets for HIV infection. It would be especially interesting to understand why the target cell population in rebound is so small, but infection was able to restart nonetheless.

The estimated parameters also lead to some interesting observations about the basic reproductive ratio R_0. Recall that threshold value for R_0 is 1: $R_0 > 1$ implies that infection will

take off, and $R_0 < 1$ means that infection cannot. In the rebound data, where there was a greater amount of confidence in the parameter estimates that the fitting algorithm converged upon, there was a mean R_0 of about 7.5. his shows that secondary infection is still far from being controlled by the immune system. This means that the bar is set quite high for the efficacy of any treatment that would have the potential to prevent rebound infection from occuring post cessation of treatment. Similarly, the early viral growth rate, r_0, is quite large during rebound, indicating of the importance of frequent sampling early on in order to capture the initial viral growth period in granular detail. Note that because the infected cell death rate (a) is relatively constant between individuals, and because of the relationship R_0 between r_0 (formula $r_0 = (R_0 - 1) * a$), their values are very closely correlated in this subject pool. When looking at the acute data, we had difficulty fitting the initial slope of exponential viral growth due to infrequent sampling. Because in many individuals the algorithm is estimating this growth to be extremely steep (very high r_0), the inaccuracy in estimating this initial growth rate gets translated into an astronomical value for R_0 (Table 5) that almost certainly cannot be trusted.

The major limitations of this study revolve around the inability to find parameter estimates for the acute data that we could be confident in. The solutions that the fitting algorithm converged upon for the acute data appeared to fall into two major categories: one where the initial slope was way too steep and very uncertain, and one where equilibrium had an unrealistically high amplitude of oscillation. Both problematic in terms of accurately fitting the viral load data. The root of this issue is likely due to the acute data not being sampled nearly as frequently as it needed to be, particularly early on in the infection. In addition to the sparsity of data points and thus an insufficient amount of information when data is collected just every

seven days, there also appeared to be issues resulting from inability to fit data with a large difference between peak viral load and set point equilibrium value. This points to the fact that it may be some combination of sparsity of observations and underlying misspecification of the given model that are resulting in the poor fit. One future step to address this particular problem could be adding an additional term to include the possibility of an immune response (Burg et al 2009), which allows for larger post-peak drops in viral load.

Beyond the inability to accurately estimate parameter values in the acute infection, there were other limitations in this study that prevented further analyses. For instance, since only two subjects were ultimately cured, it was impossible to implement any robust statistical analysis to determine whether or not parameters differed between individuals who were ultimately cured, and those who were not. It became clear throughout this analysis that the sample size simply was not large enough to make the kinds of data driven conclusions that were initially within the scope of this project. In order to truly utilize the methods developed in this thesis for parameter estimation to ultimately understanding HIV infection and treatment dynamics it is clear that a larger data set would be necessary. Though not without logistical obstacles, a study with many more individuals could allow for more data driven analysis such as clustering collections of parameters that are associated with being cured, and utilizing biological measurements to predict particular simple or composite parameter values. Such analysis could provide a more substantial link between the mathematical understanding of HIV dynamics, and how these concepts impact the ability to eventually cure actual individuals suffering from HIV.

References

Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, USA; 1991.

Borducchi EN, Cabral C, Stephenson KE, et al. Ad26/MVA therapeutic vaccination with TLR7 stimulation in SIV-infected rhesus monkeys. *Nature*. 2016;540(7632):284-287. doi:10.1038/nature20583

Box, M. J. "A Comparison of Several Current Optimization Methods, and the Use of Transformations in Constrained Problems." The Computer Journal, vol. 9, no. 1, 1966, pp. 67–77. Doi:http://dx.doi.org/10.1093/comjnl/9.1.67

Burg, Rong, Neumann, and Dahari. "Mathematical Modeling of Viral Kinetics under Immune Control during Primary HIV-1 Infection." Journal of Theoretical Biology 259, no. 4 (2009): 751-59. Doi: 10.1016/j.jtbi.2009.04.010

Cummins, Nathan W, and Andrew D Badley. "Making Sense of How HIV Kills Infected CD4 T Cells: Implications for HIV Cure." *Molecular and cellular therapies* 2 (2014): 20. *PMC*. Web. 29 Mar. 2018. Doi: 10.1186/2052-8426-2-20

Broder S. Twenty-Five Years of Translational Medicine in Antiretroviral Therapy: Promises to Keep. *Science Translational Medicine*. 2010;2(39):39ps33-39ps33. doi:10.1126/scitranslmed.3000749

Clapham HE, Tricou V, Van Vinh Chau N, Simmons CP, Ferguson NM. Within-host viral dynamics of dengue serotype 1 infection. *J R Soc Interface*. 2014;11(96). doi:10.1098/rsif.2014.0094

Chatterjee A, Smith PF, Perelson AS. Hepatitis C Viral Kinetics: The Past, Present, and Future. *Clinics in Liver Disease*. 2013;17(1):13-26. doi:10.1016/j.cld.2012.09.003

Chen HY, Mascio MD, Perelson AS, Ho DD, Zhang L. Determination of virus burst size in vivo using a single-cycle SIV in rhesus macaques. *PNAS*. 2007;104(48):19079-19084. doi:10.1073/pnas.0707449104

Dahari H, Shudo E, Ribeiro RM, Perelson AS. Mathematical modeling of HCV infection and treatment. *Methods Mol Biol*. 2009;510:439-453. doi:10.1007/978-1-59745-394-3_33

Deeks SG, Lewin SR, Ross AL, et al. International AIDS Society global scientific strategy: towards an HIV cure 2016. *Nat Med*. 2016;22(8):839-850. doi:10.1038/nm.4108

Boer RJD, Mohri H, Ho DD, Perelson AS. Turnover Rates of B Cells, T Cells, and NK Cells in Simian Immunodeficiency Virus-Infected and Uninfected Rhesus Macaques. The Journal of Immunology. 2003;170: 2479–2487. doi:10.4049/jimmunol.170.5.2479

Elemans, M. et al. Why don't CD8+ T cells reduce the lifespan of SIV-infected cells in vivo? PLOS Comput. Biol. 7, e1002200. 2011

Eisele E, Siliciano RF. Redefining the viral reservoirs that prevent HIV-1 eradication. *Immunity*. 2012;37(3):377-388. doi:10.1016/j.immuni.2012.08.010

Hill AL. Mathematical Models of HIV Latency. In: Silvestri G, Lichterfeld M, eds. *HIV Latency*. Current Topics in Microbiology and Immunology. Springer, Berlin, Heidelberg; 2017:1-26. doi:10.1007/82_2017_77

Ho DD, Neumann AU, Perelson AS, et al. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*. 1995;373(6510):123–126.

Jansson J, Wilson DP, Carr A, Petoumenos K, Boyd MA. Currently available medications in resource-rich settings may not be sufficient for lifelong treatment of HIV. *AIDS*. December 2012. doi:10.1097/QAD.0b013e32835e163d

Larder BA, Darby G, Richman DD. HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. *Science*. 1989;243(4899):1731-1734. doi:10.1126/science.2467383

Nowak MA, May RMC. *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford University Press, USA; 2000.

McLean AR, Nowak MA. Competition between zidovudine-sensitive and zidovudine-resistant strains of HIV. *AIDS*. 1992;6(1):71.

Murillo LN, Murillo MS, Perelson AS. Towards multiscale modeling of influenza infection. *Journal of Theoretical Biology*. 2013;332:267-290. doi:10.1016/j.jtbi.2013.03.024
Perelson AS, Ribeiro RM. Modeling the within-host dynamics of HIV infection. *BMC Biology*. 2013;11(1):96. doi:10.1186/1741-7007-11-96

Newville, M., Stensitzki, T., Allen, D. B., & Ingargiola, A. (2014, September 21). LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. Zenodo. Doi: http://doi.org/10.5281/zenodo.11813

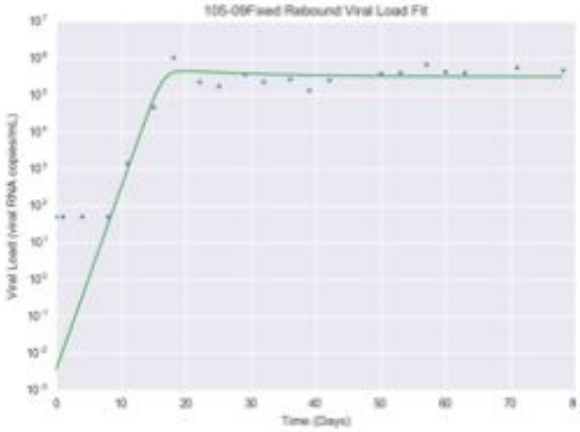
Sharp, Paul M., and Beatrice H. Hahn. "Origins of HIV and the AIDS Pandemic." *Cold Spring Harbor Perspectives in Medicine:* 1.1 (2011): a006841. *PMC*. Web. 29 Mar. 2018.

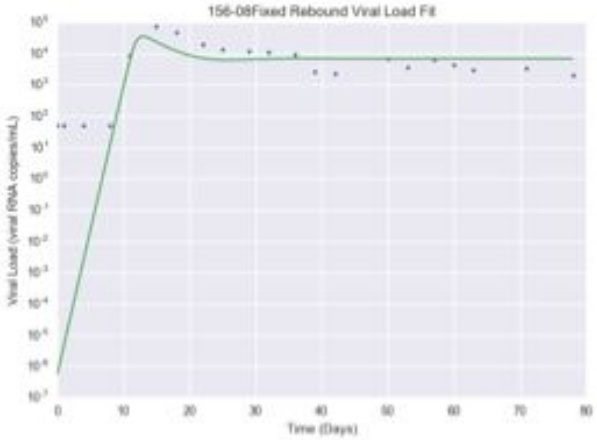
Ramratnam B, Bonhoeffer S, Binley J, et al. Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *The Lancet*. 1999;354(9192):1782-1785. doi:10.1016/S0140-6736(99)02035-8

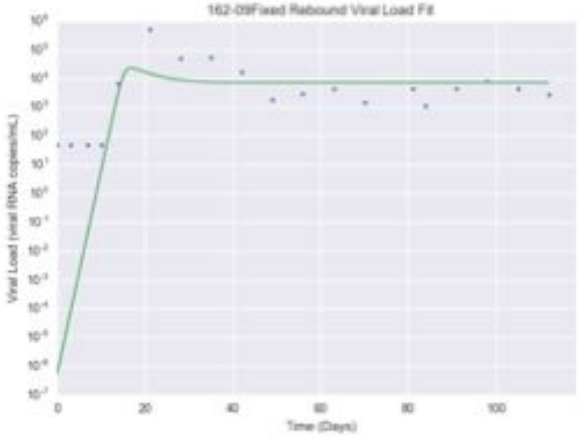
UNAIDS. *Fact Sheet - Latest Statistics on the Status of the AIDS Epidemic*.; 2017. http://www.unaids.org/en/resources/fact-sheet. Accessed February 1, 2018.

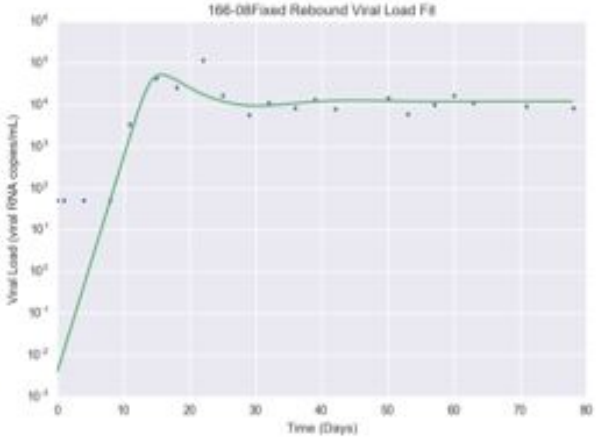
Appendix

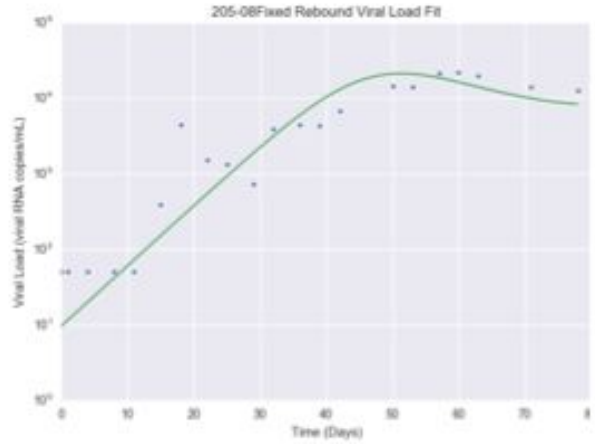
A. Final rebound fits

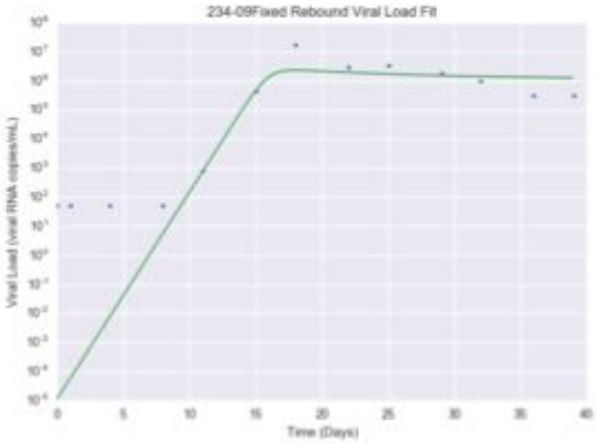


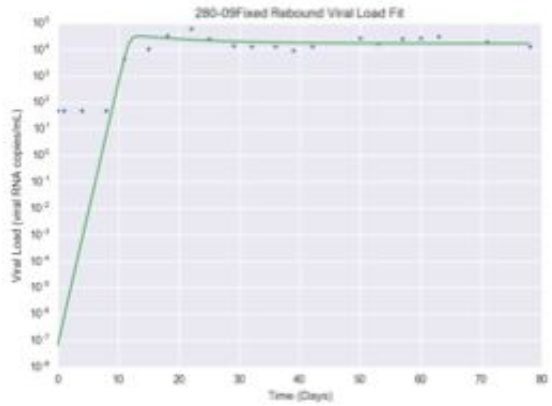


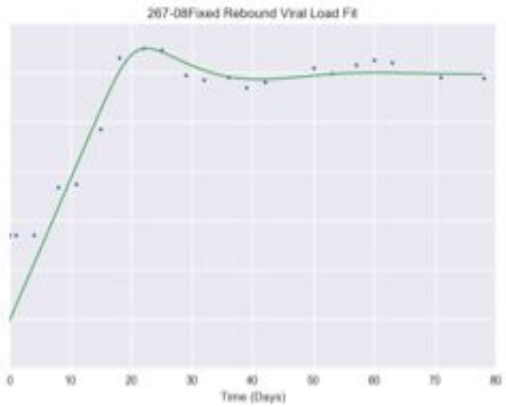


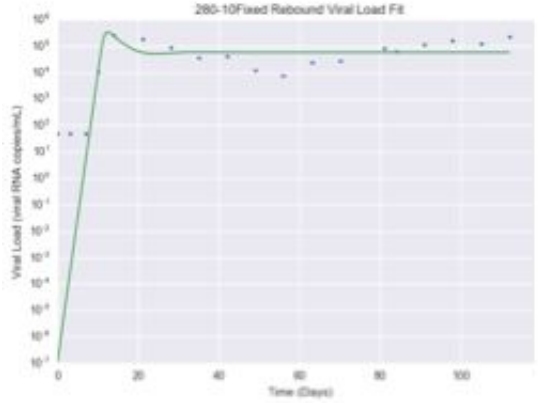


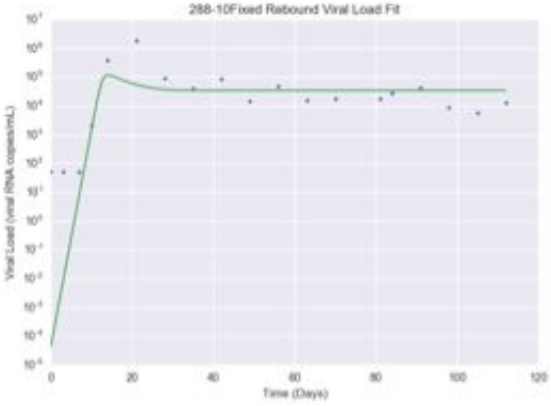


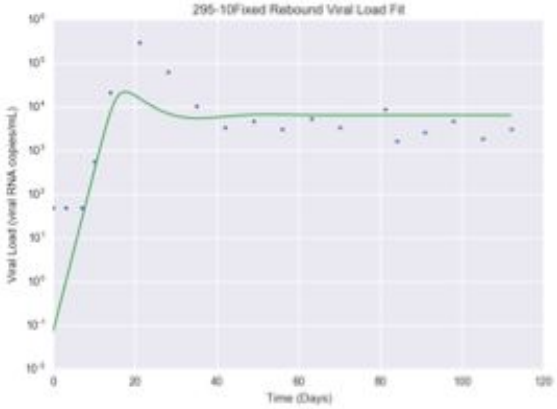


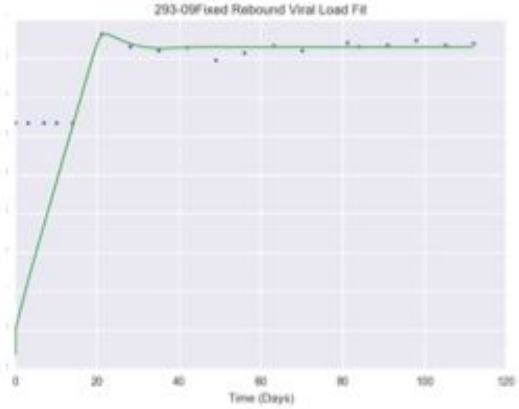

280-09Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

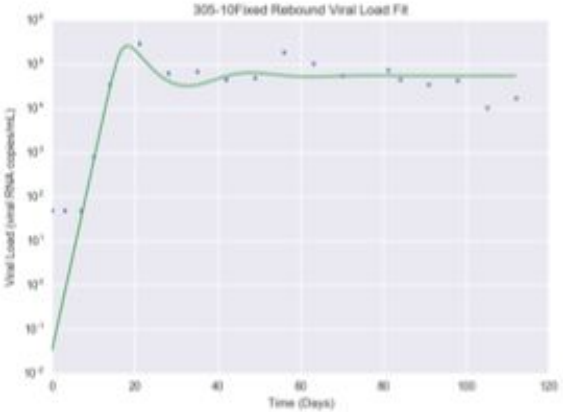

267-08Fixed Rebound Viral Load Fit
Time (Days)

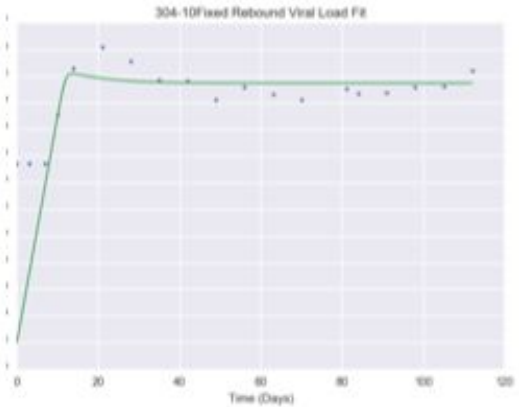

280-10Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

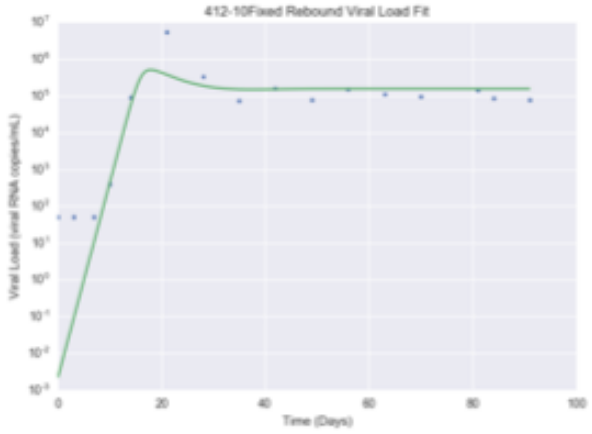

288-10Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

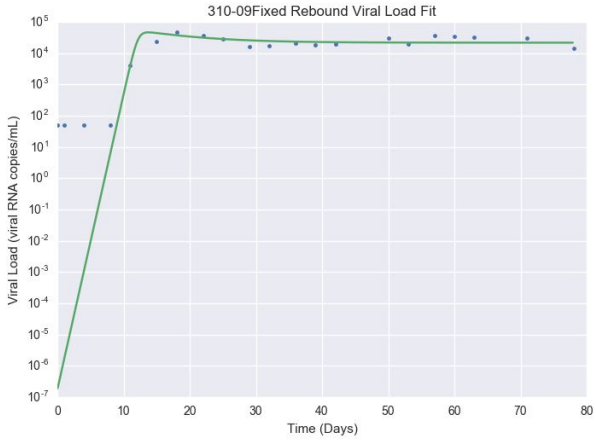

295-10Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

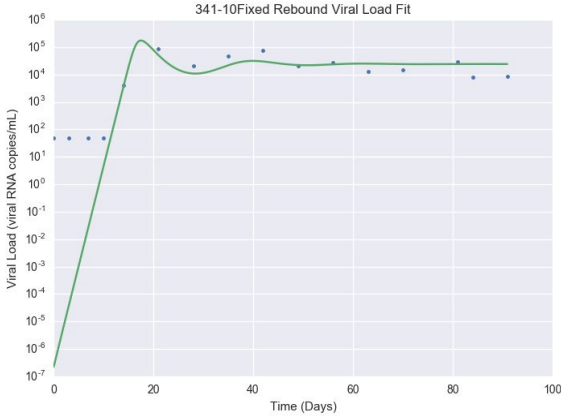

293-09Fixed Rebound Viral Load Fit
Time (Days)

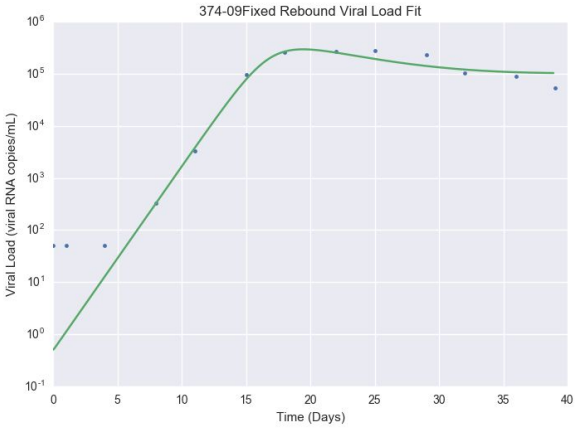

305-10Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

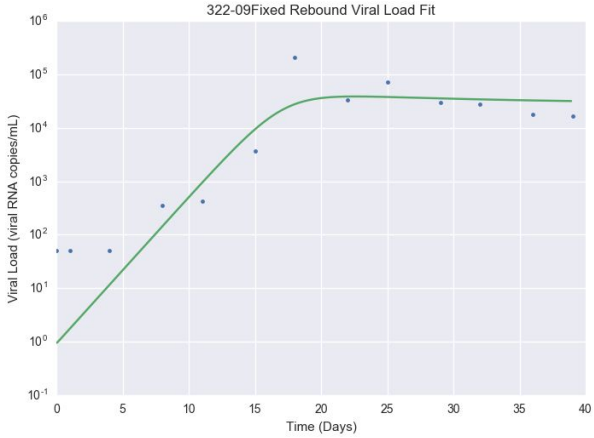

304-10Fixed Rebound Viral Load Fit
Time (Days)


412-10Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

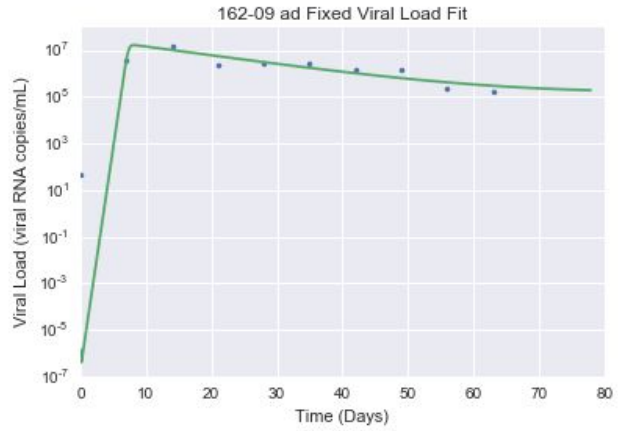

310-09Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

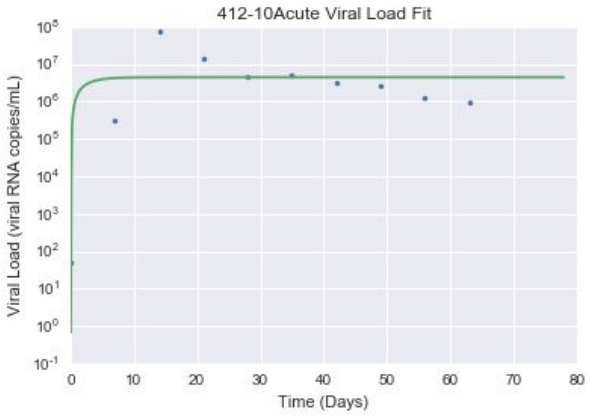

341-10Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

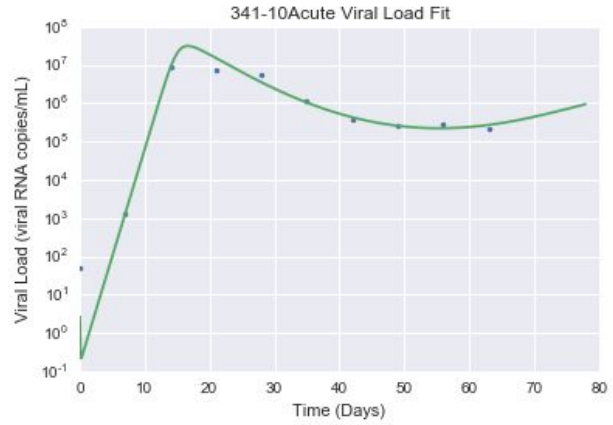

374-09Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

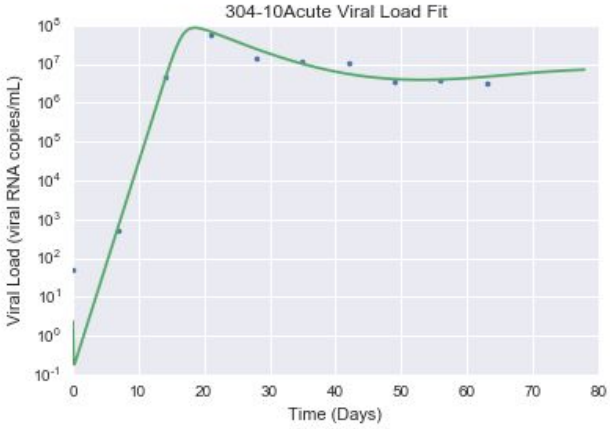

322-09Fixed Rebound Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

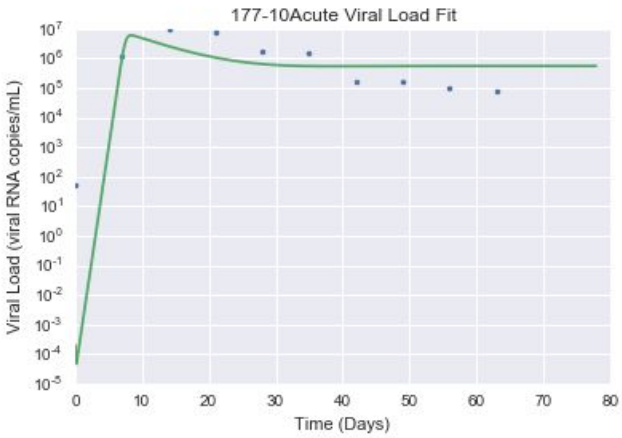
B. Final Acute Results

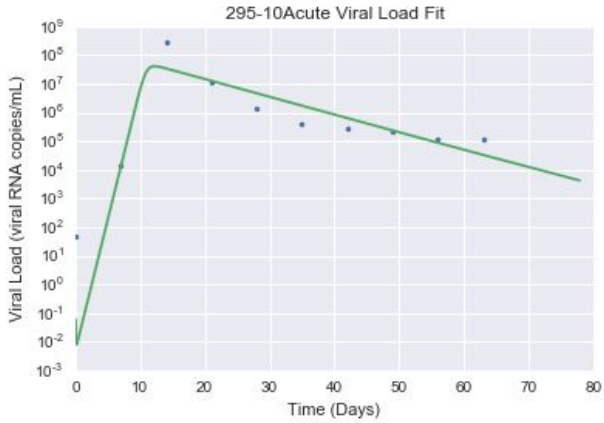


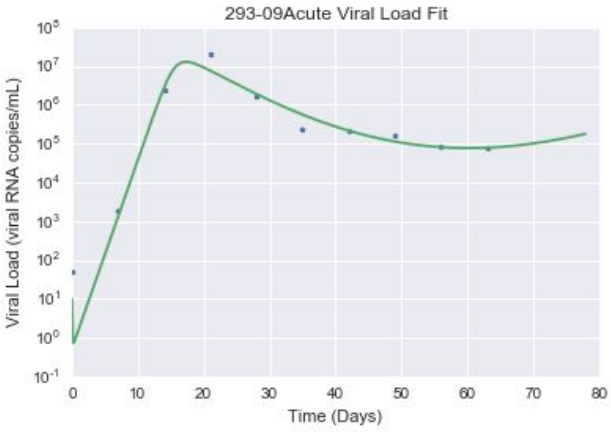


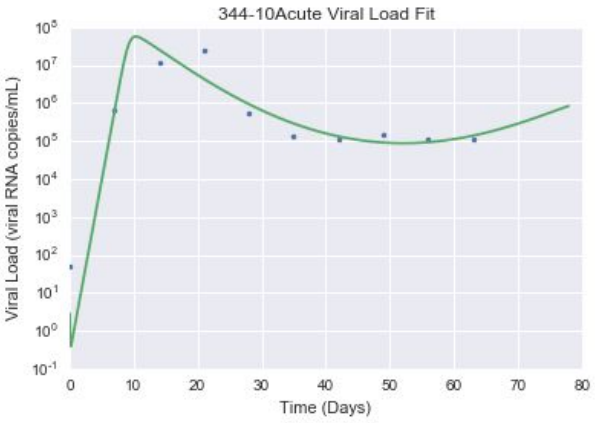


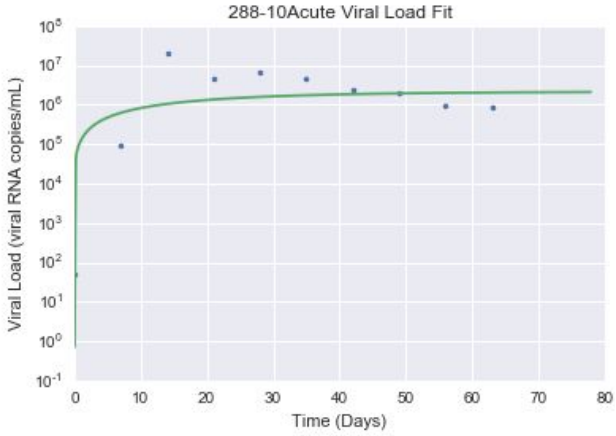


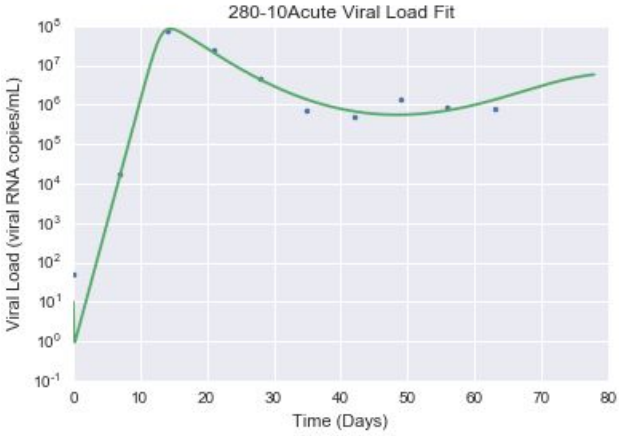


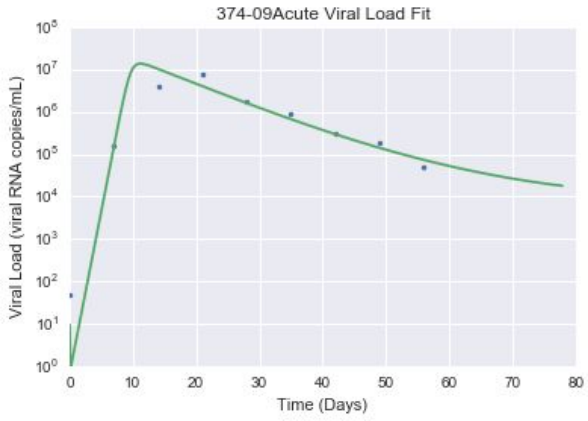


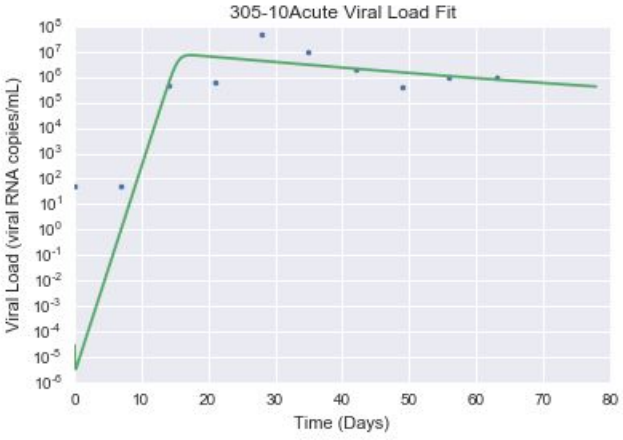

293-09Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

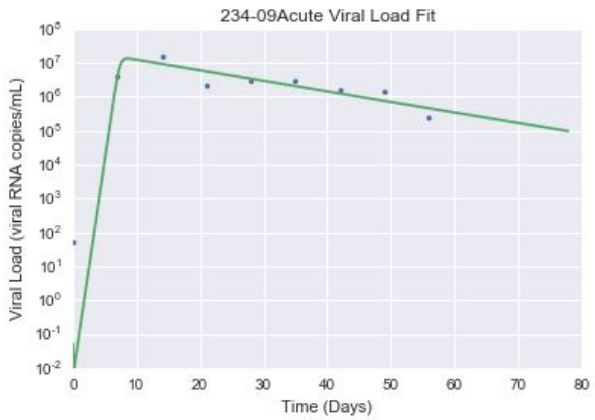

344-10Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

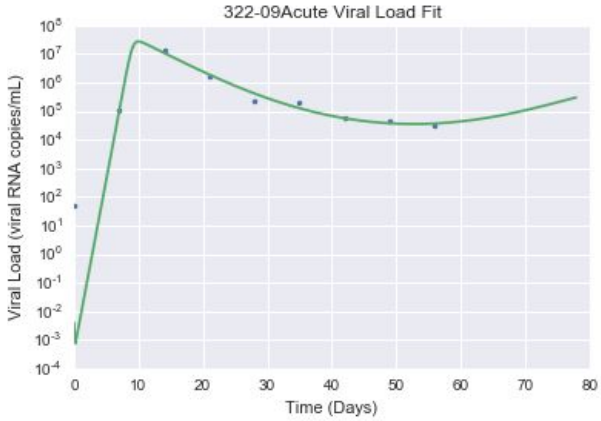

288-10Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

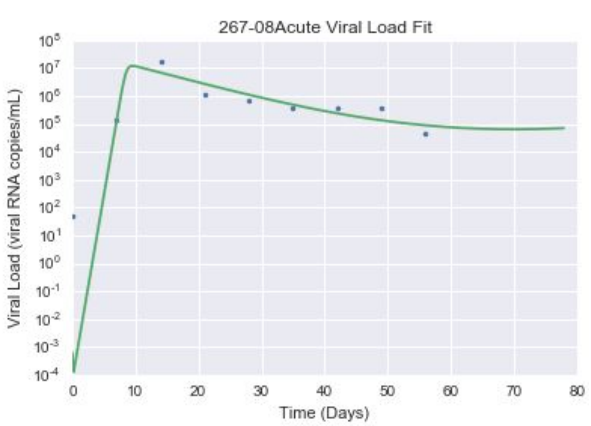

280-10Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

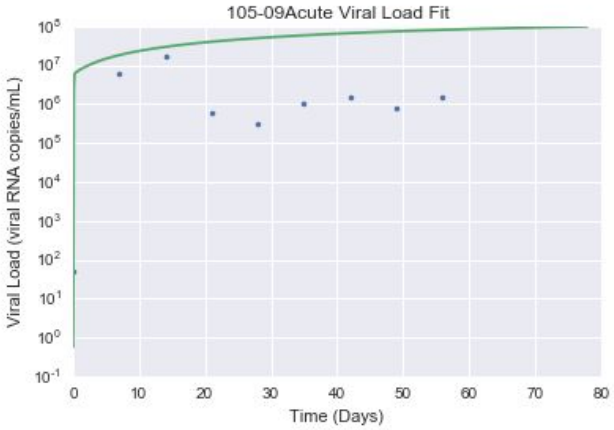

374-09Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

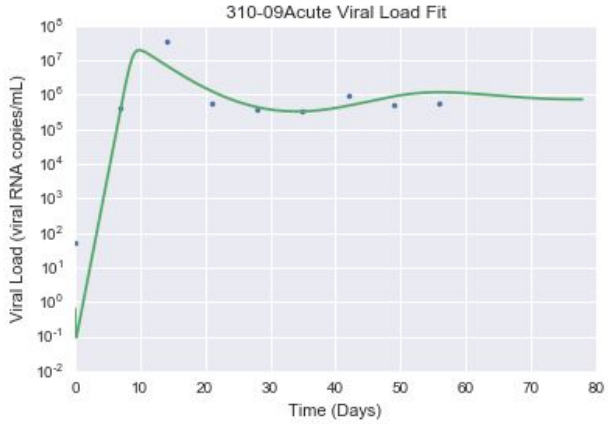

305-10Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

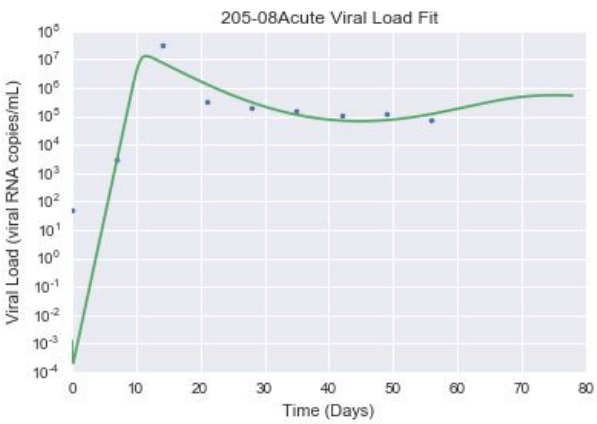

234-09Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

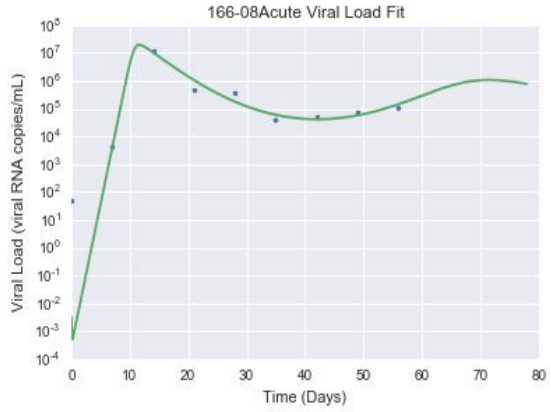

322-09Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)

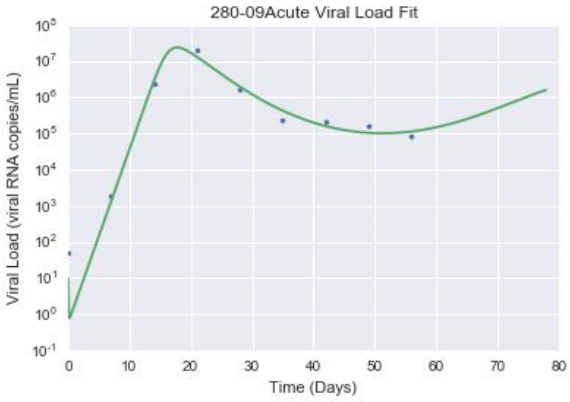

267-08Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)


105-09Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)


310-09Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)


205-08Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)


166-08Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)


280-09Acute Viral Load Fit
Viral Load (viral RNA copies/mL)
Time (Days)