A Simple Model for Human Immunodeficiency Virus Based On Erlang's Method of Stages

Samuel Swanson
Department of Mathematics, University of Florida
samueljswanson@ufl.edu

Abstract

The purpose of this paper is to select a model for HIV that uses few parameters while fitting the world prevalence and death data well. Here we consider a set of models based on Erlang's method of stages, including some with and some without social distancing. The use of stages is supported by biological studies which suggest that HIV passes through stages in each individual, although the exact number is not known. This set of models can represent such stages using a successive number of classes. To perform model selection, we compute $\mathcal{R}_0$ and use it to estimate initial values of the parameters in this model. We run thousands of iterations of a Nelder-Mead simplex search algorithm to determine the optimal values of parameters for each model and the error associated with each model. These errors are used to compute AIC$_c$ values and then the AIC$_c$ values are compared to select the most likely model. The selected model from this experiment contains the social distancing term as well as four infected classes/stages. We then perform identifiability analysis and determine that the "true values" of the parameters for this model are uniquely determinable based on the data points.

1 Introduction

Acquired Immune Deficiency Syndrome (AIDS) is a disease caused by HIV after a long incubation period (on average 8-10 years) [18]. AIDS destroys the body's ability to fight other infections through the immune system. The virus has a high virulence and mutation rate which make it particularly dangerous [14]. The disease not only has devastating health effects but also economic impacts on affected individuals, families, communities, and entire nations. Stigma causes social distancing measures between infected and uninfected individuals to occur [10]. These include precautions such as avoiding/refusing sexual contact with infected persons or using condoms. Data on cases is available from the World Health Organization, including data on the worldwide prevalence of HIV and the number of deaths each year due to AIDS.

Mathematical models are often fit to data in order to be validated. Models can enrich our knowledge of epidemiological processes on the population scale. Some existing models of HIV, such as that of Bozkurt and Peker, include an HIV negative class, an HIV positive class who know their status, and an HIV positive class who do not know their status [2]. Low-Beer and Stoneburner created an age- and sex- structured HIV epidemiological model [9]. Other examples of HIV models can be found in literature, such as in the papers of Bhunu et. al. [1] and Nyabadza et. al. [15,16]. However, many of these models use a large number of classes and parameters to model the data.

The purpose of this paper is to find a very simple model that fits the data using few parameters. We expect that the selected model can be used as the baseline for more complex and realistic models. Section 2 of this paper introduces the model set. The models we are considering have an exponential social distancing term and various numbers of stages of the infection using a technique based on Erlang's method of stages [7]. Each model is fitted to the data points using an optimization algorithm to determine the parameters that minimize the error. Model selection is then performed in order to choose the model with the highest relative likelihood out of the list of those being tested. In Section 3, identifiability analysis is performed in order to determine if the parameters are practically identifiable such that they can be uniquely determined based on the data points. Section 4 contains a discussion of the findings and Section 5 contains acknowledgments.

2 Model Selection

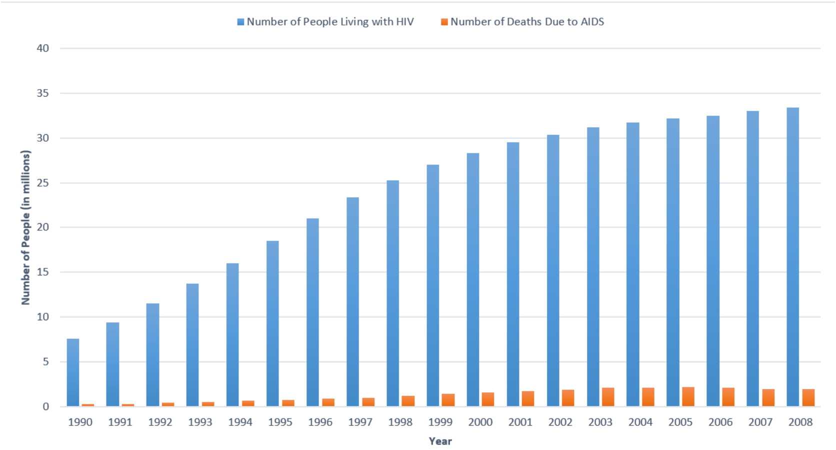

Figure 1: Prevalence of HIV and deaths due to AIDS worldwide using data from the World Health Organization website [20].

The goal of this section is to select a simple model that fits the HIV prevalence data as well as the AIDS death data (Figure 1). The following ordinary differential equation model is widely used as a baseline model for HIV:

$$\mathrm{M}_1 : \begin{cases} S'(t) = \Lambda - \beta \frac{SI}{N} - \mu S, \\ I'(t) = \beta \frac{SI}{N} - (\gamma + \mu) I, \end{cases} \tag{1}$$

In this model, $S(t)$ is the number of susceptible individuals at time t, $I(t)$ is the number of infectious individuals at time t. Susceptible individuals are recruited to the total population at a rate Λ. Susceptible individuals become infected at a transmission rate of β. Those infected become removed from the population at a disease-induced death rate of γ, which in this model is the reciprocal of the average lifespan of a person infected with HIV/AIDS. Individuals in all classes die at a natural mortality rate μ, which varies based on the population. We will fix the value of μ in all our models at the reciprocal of the average lifespan worldwide, $\frac{1}{70}$. The equation of the change in total population size is

$$N'(t) = \Lambda - \mu N - \gamma I$$

In the absence of disease at equilibrium, we have

$$N = \frac{\Lambda}{\mu}$$

Hence, we can use the total population size, N, to estimate a value for Λ. The total population size in 1990 was approximately 5.283×10^9 and in 2010 was 6.857×10^9 [19]. From this we generate an estimated value of 100 million individuals recruited per year and fix Λ at this value for all our models.

The average lifespan after infection, $\frac{1}{\gamma}$, can range anywhere between 1 to 10 or more years and has increased significantly in recent years [13]. For the parameter γ we select a randomized initial value from the interval $[\frac{1}{10}, 1]$ and then update this parameter as we fit the model.

In order to select initial values for the parameter β, we will first determine the parameter's relationship with the basic reproduction number, $\mathcal{R}_0$. Using the next-generation matrix approach, the linearized system for the infected compartments can be rewritten as

$$x' = (F - V)x$$

where

$$F = \left(\begin{array}{c} \beta \frac{S}{N} \end{array} \right), V = \left(\begin{array}{c} \gamma + \mu \end{array} \right).$$

When the system is in disease-free equilibrium $I = 0$ and $S = N$, we obtain the next generation matrix

$$K = FV^{-1} = \left(\begin{array}{c} \frac{\beta}{\gamma+\mu} \end{array} \right).$$

K has principal eigenvalue $\frac{\beta}{\gamma+\mu}$ which is the expression for $\mathcal{R}_0$ for the first model. The basic reproduction number, $\mathcal{R}_0$, for HIV can range from approximately 2 to 5 around the globe [22]. We select a randomized value for $\mathcal{R}_0$ in the interval [2,5]. With this value as well as the fixed value for μ and our randomized initial value for γ, we use the formula for $\mathcal{R}_0$ to generate an initial value for β. We then fit model M_1 to the data using an optimization algorithm to determine the fitted values for these γ and β. The algorithm we use is a variation of the Nelder-Mead simplex search method implemented in Matlab. We run this algorithm for approximately 10000 iterations for each model or until the Sum of Square Errors (SSE) value stops decreasing. We repeat these steps with 30 other sets of randomized initial values for the parameters γ and β to ensure that the algorithm converges to a global minimum rather than just a local minimum. We will use these same steps for all candidate models.

Model M_1 does not fit the data. This can be seen from the high SSE of 696.60 after fitting the data to this model, as well as the poor fit on the graph in Figure 2.

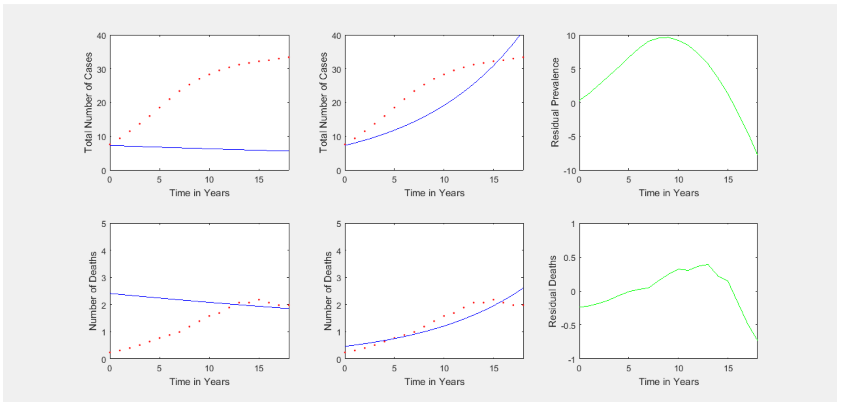

Figure 2: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M_1. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit. This model clearly does not fit the data.

Our goal is to find a simple model that more accurately represents the data. For the following models, we add an exponential social distancing term, $e^{-\kappa \frac{I}{N}}$, where κ is the rate of social distancing. A similar exponential term is seen in a paper by Williams, where it is used to allow for heterogeneity in the risk of infection among people. The paper asserts that high risk people are likely to be infected before low risk people so the transmission parameter decreases as prevalence increases [21]. We introduce multiple models with this social distancing term for HIV, and the parameters for these models are summarized in Table 1. The interplay between human behavior and infectious disease has been extensively studied through models in [10], showing the ability for a social distancing effect to reduce the transmissibility of a disease.

Table 1: List of parameters in HIV models

Notation	Meaning	Units	Value
Λ	recruitment rate of susceptible individuals	$\frac{people}{year}$	100 (million)
β	transmission rate for HIV	$\frac{1}{year}$	as fitted
κ	coefficient of social distancing	unitless	as fitted
μ	natural mortality rate	$\frac{1}{year}$	$\frac{1}{70}$
γ	disease-induced death rate	$\frac{1}{year}$	as fitted

We now introduce Model M$_2$, which is the same as Model M$_1$ except with a social-distancing term. Model M$_2$ also has $\mathcal{R}_0 = \frac{\beta}{\gamma+\mu}$

$$\mathrm{M}_2 : \begin{cases} S'(t) = \Lambda - \beta\frac{SI}{N}e^{-\kappa\frac{I}{N}} - \mu S, \\ I'(t) = \beta\frac{SI}{N}e^{-\kappa\frac{I}{N}} - (\gamma+\mu)I, \end{cases} \tag{2}$$

We fit model M_2 to the data and present the initial and final fit and residuals in Figure 3.

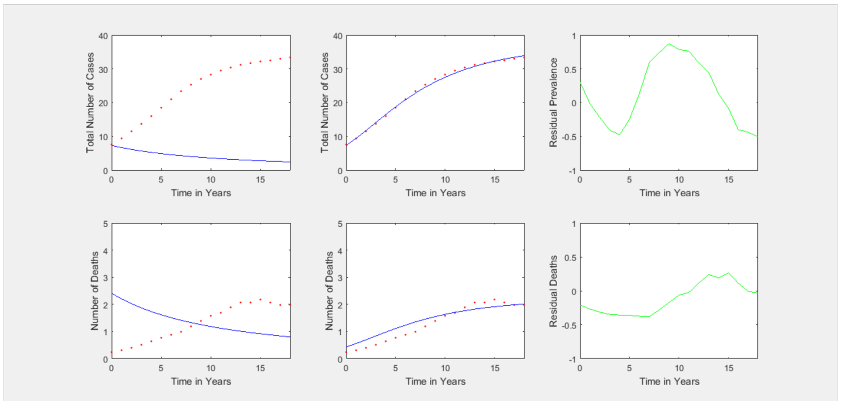

Figure 3: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M$_2$. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit.

The removal rate in model M$_2$ is exponentially distributed, which may not be entirely realistic. Each of the following gamma-models are based on Erlang's method of stages which allows model exit rates which are non-exponentially distributed [3]. This approach is typically applied to stochastic HIV models; however, we use a deterministic variant that uses a variable number of stages with the durations of stay in each stage being independent and identically distributed exponential variables [12]. Researchers have determined that HIV passes through various stages, the exact number of which is not known [5]. Increasing the number of sequential infected classes in this model may help us to understand what the optimal number of stages is. For the following models, γ is not only the disease-induced death rate but also the rate of progression from one stage of the disease to the next. For the next model, M$_3$, $I = I_1 + I_2$.

$$\mathrm{M}_3 : \begin{cases} S'(t) = \Lambda - \beta\frac{SI}{N}e^{-\kappa\frac{I}{N}} - \mu S, \\ I_1'(t) = \beta\frac{SI}{N}e^{-\kappa\frac{I}{N}} - (\gamma+\mu)I_1, \\ I_2'(t) = \gamma I_1 - (\gamma+\mu)I_2, \end{cases} \tag{3}$$

We fit model M_3 to the data and present the fit and residuals in Figure 4.

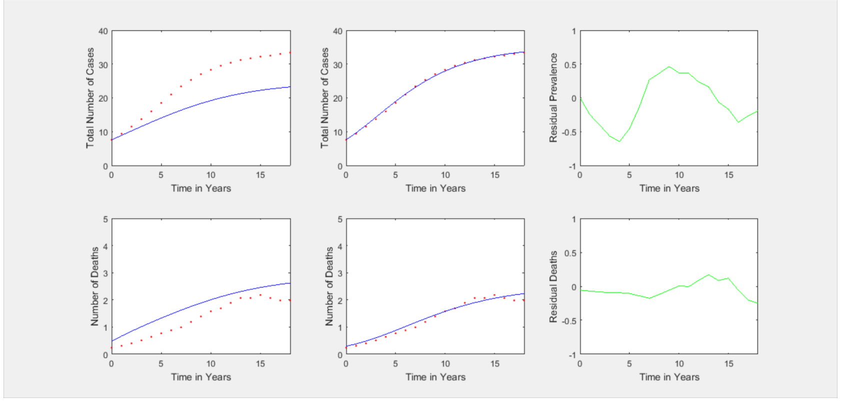

Figure 4: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M_3. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit.

For the next model, $I = I_1 + I_2 + I_3$.

$$\mathrm{M}_4 : \begin{cases} S'(t) = \Lambda - \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - \mu S, \\ I_1'(t) = \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - (\gamma + \mu) I_1, \\ I_2'(t) = \gamma I_1 - (\gamma + \mu) I_2, \\ I_3'(t) = \gamma I_2 - (\gamma + \mu) I_3, \end{cases} \tag{4}$$

We fit model M_4 to the data and present the fit and residuals in Figure 5.

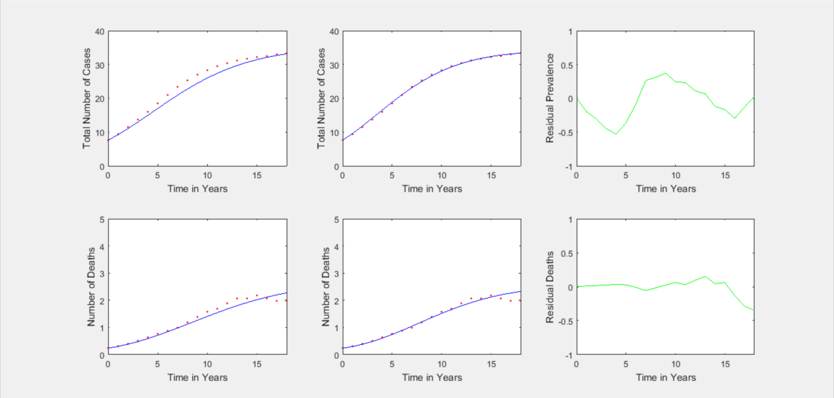

Figure 5: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M_4. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit.

For the next model, $I = I_1 + I_2 + I_3 + I_4$.

$$\mathrm{M}_5 : \begin{cases} S'(t) = \Lambda - \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - \mu S, \\ I_1'(t) = \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - (\gamma + \mu) I_1, \\ I_2'(t) = \gamma I_1 - (\gamma + \mu) I_2, \\ I_3'(t) = \gamma I_2 - (\gamma + \mu) I_3, \\ I_4'(t) = \gamma I_3 - (\gamma + \mu) I_4, \end{cases} \tag{5}$$

We fit model M_5 to the data and present the fit and residuals in Figure 6.

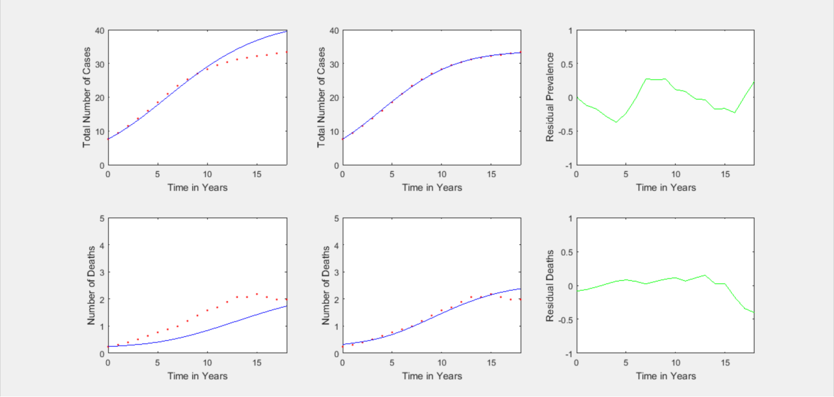

Figure 6: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M_5. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit.

For the next model, $I = I_1 + I_2 + I_3 + I_4 + I_5$.

$$\mathrm{M}_6 : \begin{cases} S'(t) = \Lambda - \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - \mu S, \\ I_1'(t) = \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - (\gamma + \mu) I_1, \\ I_2'(t) = \gamma I_1 - (\gamma + \mu) I_2, \\ I_3'(t) = \gamma I_2 - (\gamma + \mu) I_3, \\ I_4'(t) = \gamma I_3 - (\gamma + \mu) I_4, \\ I_5'(t) = \gamma I_4 - (\gamma + \mu) I_5, \end{cases} \tag{6}$$

We fit model M_6 to the data and present the fit and the residuals in Figure 7. Since the SSE has increased from that of Model M_5, we will stop the process of adding more stages at this point.

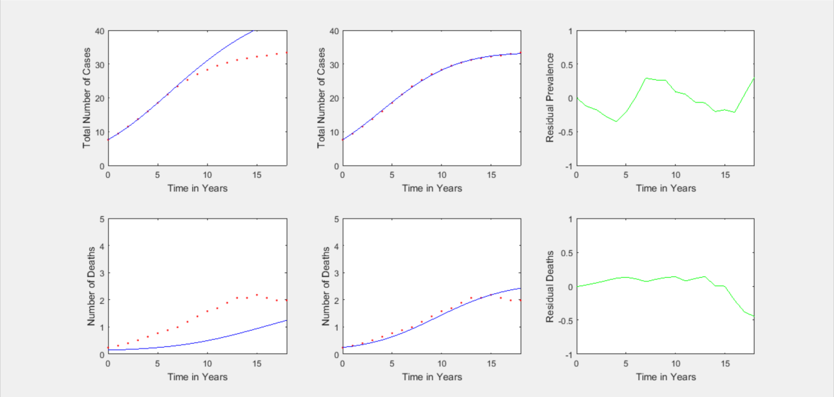

Figure 7: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M$_6$. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit.

In the previous models, the prevalence data is fitted very well; however, the death data does not fit quite as well. Thus, for the next model, we will modify Model M_5 such that there is a separate parameter for deaths, changing the final γ to γ_2, in order to determine if this will improve the fitting of the death data. As in Model M_5, $I = I_1 + I_2 + I_3 + I_4$.

$$\mathrm{M}_7: \begin{cases} S'(t) = \Lambda - \beta\frac{SI}{N}e^{-\kappa\frac{I}{N}} - \mu S, \\ I_1'(t) = \beta\frac{SI}{N}e^{-\kappa\frac{I}{N}} - (\gamma+\mu)I_1, \\ I_2'(t) = \gamma I_1 - (\gamma+\mu)I_2, \\ I_3'(t) = \gamma I_2 - (\gamma+\mu)I_3, \\ I_4'(t) = \gamma I_3 - (\gamma_2+\mu)I_4, \end{cases} \tag{7}$$

We fit model M_7 to the data and present the fit and the residuals in Figure 9.

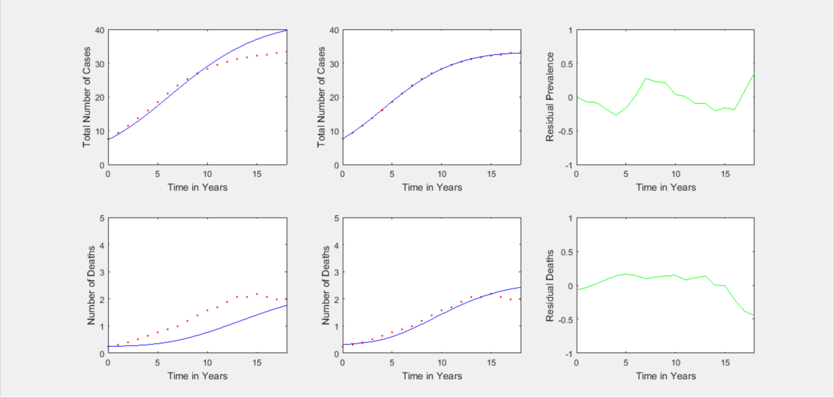

Figure 8: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M_7. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit.

We will also repeat this with a separate parameter for each of the transitions between classes in Model M_8. Again, $I = I_1 + I_2 + I_3 + I_4$.

$$\mathrm{M}_8: \begin{cases} S'(t) = \Lambda - \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - \mu S, \\ I_1'(t) = \beta \frac{SI}{N} e^{-\kappa \frac{I}{N}} - (\gamma + \mu) I_1, \\ I_2'(t) = \gamma I_1 - (\gamma_2 + \mu) I_2, \\ I_3'(t) = \gamma_2 I_2 - (\gamma_3 + \mu) I_3, \\ I_4'(t) = \gamma_3 I_3 - (\gamma_4 + \mu) I_4, \end{cases} \tag{8}$$

We fit model M_8 to the data and present the fit and the residuals in Figure 10.

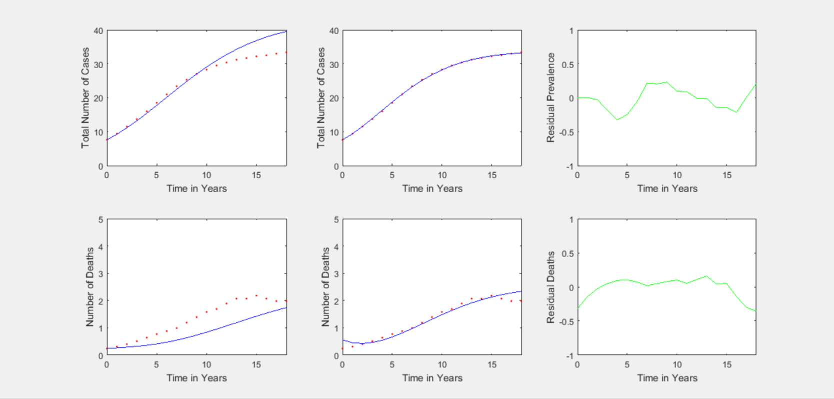

Figure 9: The top row contains prevalence fittings and the bottom row contains deaths fittings. The red dots are the data points and the blue curves are the models. In the left column are the plots using the initial guess of parameters for model M_8. In the middle column are the solution curves using parameters after fitting. In the right column are plots of the residuals left over from the fit.

Lastly, we include Model M$_9$ which has 4 infected stages like Model M$_5$, but without the exponential social distancing term. Again, $I = I_1 + I_2 + I_3 + I_4$.

$$\mathrm{M}_9 : \begin{cases} S'(t) = \Lambda - \beta\frac{SI}{N} - \mu S, \\ I_1'(t) = \beta\frac{SI}{N} - (\gamma + \mu)I_1, \\ I_2'(t) = \gamma I_1 - (\gamma + \mu)I_2, \\ I_3'(t) = \gamma I_2 - (\gamma + \mu)I_3, \\ I_4'(t) = \gamma I_3 - (\gamma + \mu)I_4, \end{cases} \tag{9}$$

This model does not fit the prevalence or death data, which can be seen through it's extremely high SSE of 618.80 (See Table 3). We conclude that the presence of an exponential social distancing term is essential for HIV models to fit the data.

Table 2 gives the fitted values of the parameters for each of the models M_1 through M_9.

Model	κ	β	γ	γ_2	γ_3	γ_4
1	-	0.1741	0.0633	-	-	-
2	362.40	0.5051	0.0593	-	-	-
3	319.15	0.4197	0.1483	-	-	-
4	287.74	0.3635	0.2362	-	-	-
5	274.39	0.3421	0.3247	-	-	-
6	264.48	0.3280	0.4141	-	-	-
7	256.15	0.3176	0.5049	0.4873	-	-
8	211.40	0.4349	0.4325	0.5456	0.5122	0.5993
9	-	0.2086	0.5461	-	-	-

Table 2: Fitted values of the parameters for each model.

We compare the accuracy of the models using the Sum of Square Errors (SSE) and the Akaike Information Criterion (AIC). The SSE is calculated as

$$SSE = \sum (y_i - \hat{y}_i)^2$$

where y_i is the value the model predicts and $\hat{y}_i$ is the value from the data set. The AIC allows us to decide which model fits best based on having both the lowest error and the least number of parameters. It is calculated as

$$AIC = 2k + n\left(\ln\left(\frac{SSE}{n}\right)\right)$$

With too many parameters, the models can become overly complex. This can lead to overfitting in which case the models describe the random noise in addition to the actual relationship, reducing the effectiveness of the model. Because of the small number of data points proportional to the number of parameters fitted, we use the AIC_c, which is corrected for finite sample sizes. The formula for AIC_c is as follows,

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

where k is the number of parameters fitted and n is sample size [4].

The AIC_c and SSE for all models are listed in Table 3 below.

Model	SSE	Param fitted	AIC_c	Δ_j	W
(1)	696.60	2	73.184	$\Delta_1 = 119.324$	5.258×10^{-27}
(2)	5.888	3	-14.659	$\Delta_2 = 31.481$	6.247×10^{-8}
(3)	2.206	3	-33.312	$\Delta_3 = 12.828$	7.016×10^{-4}
(4)	1.607	3	-39.331	$\Delta_4 = 6.809$	0.0142
(5)	1.123	3	-46.140	$\Delta_5 = 0$	0.428
(6)	1.180	3	-45.200	$\Delta_6 = 0.941$	0.2676
(7)	0.986	4	-45.355	$\Delta_7 = 0.785$	0.2892
(8)	0.944	6	-38.039	$\Delta_8 = 8.101$	0.0074
(9)	618.80	2	70.909	$\Delta_8 = 117.042$	1.627×10^{-26}

Table 3: List of SSE, AIC_c, and Relative Likelihood for Various Models

The column Δ_j is calculated as

$$\Delta_j = AIC_j - min(AIC)$$

The final column lists the relative likelihood of each model, given by

$$\frac{e^{-\Delta_j/2}}{\sum e^{-\Delta_j/2}}$$

Looking at the table, we can see that Model M$_5$ fits the data the best; however, Models M$_6$ and M$_7$ also have some support in the data.

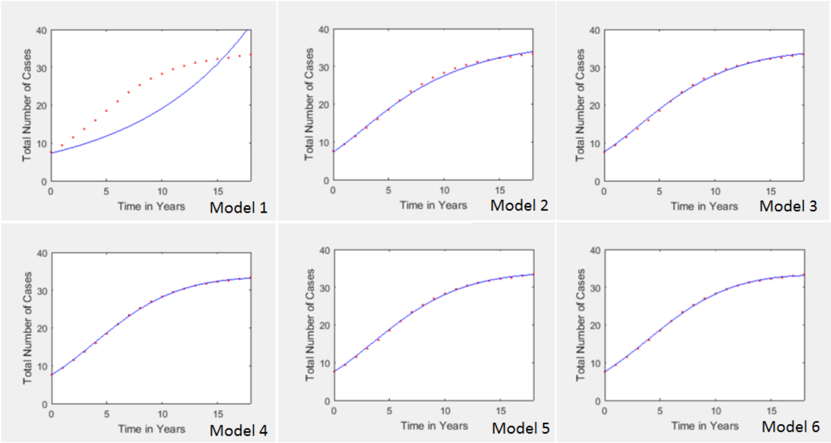

Figure 10: This figure contains the models fitted to the prevalence data. The data points are in red and the solutions to the models are in blue.

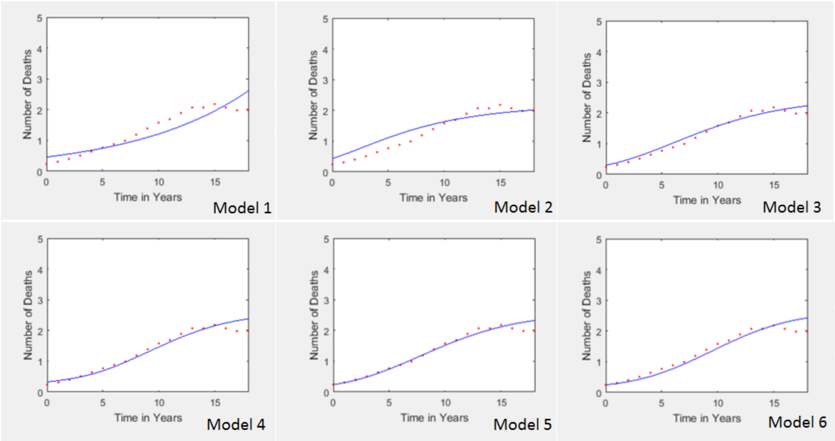

Figure 11: This figure contains the models fitted to the death data. The data points are in red and the solutions to the models are in blue.

3 Identifiability Analysis

The results of this study depend greatly on being able to estimate the parameters using the HIV prevalence and deaths data. We must determine if the parameters in the models we are using are identifiable to ensure that we are able to properly estimate the "true values" of the parameters. If multiple sets of parameters lead to the same outputs, the model is considered non-identifiable and we are unable to estimate these "true values". For Model M_1, we use a differential algebra approach in order to test its identifiability. This can be calculated by hand using a method from [11]. In this model we have observed $I(t)$, the total prevalence of the disease in year t, as well as $\gamma I(t)$, the number of deaths in year t. Thus γ is identifiable. In order to determine if β is identifiable we will obtain an input-output function. We eliminate the unobserved state, S, from the system and obtain an equation in I and its derivatives. We do so by solving for S in the first equation and then differentiating this with respect to t to obtain S'. We then replace S and S' in the first equation with these expressions. We obtain the following input-output function:

$$\Lambda - (I' + (\gamma + \mu)I) - \mu \frac{I'I + (\gamma + \mu)I^2}{\beta I - I' - (\gamma + \mu)I} = S'$$

where

$$S' = \frac{[\beta I - I' - (\gamma + \mu)I][(I')^2 + II'' + 2II'(\gamma + \mu)] - [I'I + (\gamma + \mu)I^2][\beta I' - I'' - (\gamma + \mu)I']}{[\beta I - I' - (\gamma + \mu)I]^2}.$$

In this approach, the ODE model is structurally identifiable if there is an injective map from the parameter space to the coefficients of the input-output equations. Since λ and μ are fixed in this model and we have determined that γ is identifiable, β is also identifiable. From this we conclude that all the parameters in Model M_1 are structurally identifiable as long as Λ and μ are held constant.

Structural identifiability analysis of Models M_2 through M_8 is not possible by means of the differential algebra approach because the social distancing term is exponential and this approach only works with rational functions [17]. Thus we will use Monte Carlo simulations in order to study the practical identifiability of the parameters in Model M_5. This is done as follows:

1. Estimate the parameters using the least squares method. These estimated values, which can be found in Table 2, are considered to be the true parameters $\boldsymbol{p}_0$.

2. Solve the epidemiological model numerically with the true parameter set $\boldsymbol{p}_0$ and obtain the output vector of prevalence and death values predicted by the model at the discrete data time points.

3. Using these predicted values and the actual values from the data at these time points, calculate the normalized residuals.

4. Generate M = 1000 residual vectors chosen from a normal distribution in which the mean is the residuals generated in the previous step and the standard deviation is the measurement error, σ_0, which we first set to 0.

5. Add these residual values to the actual data points to generate M = 1000 sets of simulated data.

6. Fit the model M_5 to the simulated data sets using the Nelder-Mead algorithm in MATLAB to estimate the parameter sets for the simulated data.

7. Calculate the average relative estimation error for each parameter by

$$ARE(p^{(k)}) = 100\% \frac{1}{M} \sum_{j=1}^{M} \frac{|p_0^{(k)} - p_j^{(k)}|}{p_0^{(k)}}$$

   where $p^{(k)}$ is the k^{th} parameter in the set.

8. Finally, repeat all the steps with increased levels of noise, setting σ_0 to 5% and 30%.

We report the computed ARE values of each parameter in Table 4.

Parameter	ARE 0%	ARE 5%	ARE 30%
β	$3.669 \times 10^{-5}\%$	$6.154 \times 10^{-2}\%$	$5.915 \times 10^{-1}\%$
k	$9.535 \times 10^{-4}\%$	$1.067 \times 10^{-1}\%$	1.168 %
γ	$7.152 \times 10^{-5}\%$	$7.280 \times 10^{-2}\%$	$7.096 \times 10^{-1}\%$

Table 4: Average Relative Estimation (ARE)errors from the Monte Carlo Simulations

We add noise to the model at the actual data points by drawing residuals from a normal distribution and adding them to the model output. If the model were structurally identifiable, then when there is no noise in the data, the AREs should be very close to 0 if not 0 [6]. As a strict measure, when $\sigma_0 = 5\%$, all AREs below 5% are considered identifiable. Table 4 shows that the ARE values for all the variables are relatively low, even when noise is added, so we claim that these parameters are practically identifiable.

4 Discussion

In this paper, we wish to select a simple epidemiological model of human immunodeficiency virus that fits both the prevalence data and the AIDS death data. We define a set of models with or without an exponential social distancing term and stages based on an Erlang's distribution.

The inclusion of an exponential decay term representing social distancing greatly improves the fit of the model to the data. Increasing the number of sequential infected classes improves the fit until the model reaches four infected classes in Model M_5. After this point, we see the error begins to increase. According to research studies, HIV passes through many stages within each host; however, it is not known exactly how many steps and transitions occur [3]. These classes in the model may help to represent this phenomenon. The selection of M_5 as the best fitted model suggests that on average HIV goes through 4 stages lasting approximately 3 years each. This roughly agrees with the findings of [8].

The figures with the final fits and residuals show that all of the models with the social distancing term appear to fit the prevalence data well with randomly distributed residuals. From the plots of the AIDS death models and residuals, it can be seen that the models overestimate the number of deaths in later years. This may be due to improved medications in later years that cause the death rate to be time-dependent. In Model M_7 we attempt to improve the fit of the AIDS death data by replacing the final γ which corresponds to the rate at which infected individuals are removed from the population with a separate parameter, γ_2. This does lower the error of the fit, although it does not lower it enough to justify an extra parameter being fit according to the AIC$_c$. In Model M_8 we attempt to use a separate parameter for each transition rate between stages which lowers the error but greatly increases the AIC$_c$ due to the increased number of parameters.

We make another attempt to improve the fit of the model to the death data by using the selected model, M_5, but with weighted least squares for the fitting of parameters. Because the average values for the number of deaths are approximately one tenth the magnitude of the average values for prevalence, we multiply the error from deaths by ten to bring it near the order of the error for the prevalence data. This also does not appear to improve the fit to the death data. This model clearly fits the prevalence data and fits the AIDS death data except in the last three years of data points, suggesting that a time-dependent death rate may create a better fitting and more complex model. In the future, it would be interesting to investigate a model similar to the selected model in this paper except with the disease-induced death rate, γ, being a function of time such as a step-function rather than a parameter.

Finally, we conduct practical identifiability analysis on the selected model using a Monte Carlo method and determine that all of its parameters are practically identifiable.

Overall, we were able to find a simple model that fits the prevalence data well and fits the death data well except for in the most recent years. We also discovered that this model is practically identifiable and identified the constants which could be useful in future work.

5 Acknowledgments

The author acknowledges support from NSF grant DMS-1515661. The author thanks Necibe Tuncer and Maia Martcheva for reading over this paper and making suggestions for improvement.

References

[1] C.P. Bhunu, S. Mushayabasa, H. Kojouharov, J.M. Tchuenche, Mathematical Analysis of an HIV/AIDS Model: Impact of Educational Programs and Abstinence in Sub-Saharan Africa, *J. Math. Biol.* **57**, (2008), 557-593

[2] F. Bozkurt and F. Peker, Mathematical Modelling of HIV Epidemic and Stability Analysis, *Advances in Difference Equations*, (2014)

[3] P. Bratley, B.L. Fox, L.E. Schrage, A Guide to Simulation, *Springer-Verlag*, New York, (2012)

[4] K.P.Burnham, D.R. Anderson, Model Selection and Multimodal Inference: A Practical Information-theoretic Approach, 2nd edition *Springer*, New York, (2002)

[5] P. Lemey, A. Rambaut, O. Pybus, HIV Evolutionary Dynamics Within and Among Hosts, *AIDS*, Rev. 8 (2006), 25-140.

[6] X.R. Li, Z. Zhao, Relative Error Measures for Evaluation of Estimation Algorithm, *7th International Conference on Information Fusion*, (2005)

[7] A. Lloyd, Realistic Distributions of Infectious Periods in Epidemic Models: Changing Patterns of Persistence and Dynamics, *Theoretical Population Biology* **60**, (2001), 59-71

[8] I.M. Longini et al., Statistical Analysis of the Stages of HIV Infection Using a Markov Model, *Statistics in Medicine* **8**, (1989), 831-843

[9] D. Low-Beer and R.L. Stoneburner, An Age- and Sex-Structured HIV Epidemiological Mode: Features and Applications, *Bull World Health Organ* **75**, (1997) 213-221

[10] P. Manfridi and A. D'Onofrio, Modeling the Interplay Between Human Behavior and the Spread of Infectious Disease, *Springer*, New York, (2013)

[11] G. Margaria, E. Riccomagno, M. Chappell, H. Wynn, Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences, *Mathematical Biosciences* **174**, (2001), 1-26

[12] M. Martcheva An Introduction to Mathematical Epidemiology, *Springer*, New York, (2015)

[13] NAM, http://www.aidsmap.com/Life-expectancy-now-considerably-exceeds-the-average-in-some-people-with-HIV-in-the-US/page/2816267/

[14] M. Nowak, R. May, Mathematical Biology of HIV Infections: Antigenic Variation and Diversity Threshold, *Mathematical Biosciences* **106**, (1991), 1-21

[15] F. Nyabadza et al., Analysis of an HIV/AIDS Model with Public-Health Information Campaigns and Individual Withdrawal, *J. Biol. Syst.* **18**, (2010), 357

[16] F. Nyabadza, Z. Mukandavire, Modelling HIV/AIDS in the Presence of an HIV Testing and Screening Campaign, *J. Theoret. Biol.* **280**, (2011), 167-179

[17] M.P. Saccomani, Structural vs Practical Identifiability in Systems Biology, *IWBBIO*, (2013), 305-313

[18] A.C. Tsai et al., Internalized Stigma, Social Distance, and Disclosure of HIV Seropositivity in Rural Uganda, *Ann Behav Med.* **46**, (2013)

[19] U.S. Census Bureau, Midyear Population for the World: 1950-2050,
https : //www.census.gov/population/international/data/idb/worldpoptotal.php

[20] World Health Organization, HIV/AIDS Global Epidemic Data and Statistics, *http : //www.who.int/hiv/data/global_data/en/*

[21] B. Williams, Fitting and projecting HIV epidemics: Data, structure and parsimony, (2014), *https : //arxiv.org/ftp/arxiv/papers*/1412/1412.2788*.pdf*

[22] B. Williams, E. Gouws, R0 and the elimination of HIV in Africa: Will 90-90-90 be sufficient? Available: http://arxiv.org/abs/1304.3720.