January 2021

Parameter estimation and optimization for biological mathematical models using Bayesian statistics

Renee Dale
Louisiana State University and Agricultural and Mechanical College

PARAMETER ESTIMATION AND OPTIMIZATION FOR BIOLOGICAL MATHEMATICAL MODELS USING BAYESIAN STATISTICS

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Masters in Applied Statistics

in

The Department of Experimental Statistics

by
Renee Marie Dale
BS Biology, LSU, 2013
BA Philosophy, LSU, 2013
MS Biology, LSU, 2015
PhD Biology, LSU, 2019
May 2021

Acknowledgments

I would like to thank Dr. Guo for encouraging me to stick with it, and Dr. Escobar for suggesting I continue in the initial probability class. I would also like to thank Dr. Li for serving on my committee.

Table of Contents

Acknowledgments ii

Abstract iv

Chapter 1. Bayesian Statistics and Mathematical Biology 1
1.1. Mathematical modeling in biology 1
1.2. Bayesian statistics and parameter estimation 5

Chapter 2. Estimating Epidemiological Parameters of a Stochastic Differential Model of HIV Dynamics Using a Hierarchical Bayesian Model 8
2.1. Introduction 8
2.2. Materials and methods 11
2.3. Results 15
2.4. Discussion 25

Appendix A. Rights and Permissions for Previously Published Articles 27

Appendix B. Diagnosed Population R Code 29

Appendix C. Undiagnosed Population R Code 35

References 42

Vita 46

Abstract

In the field of biology, mathematical models are increasingly used to address biological questions and the large data sets generated in experimental studies. Mathematical models traditionally are simplified and structured to be analytically tractable, but computing power allows for more complex, larger models. Bayesian statistics lends itself naturally to address parameter estimation problems in these large models. Bayesian statistical inference is utilized in this thesis to obtain parameter estimates from a sparse data set on populations in the HIV epidemic. Current estimates of the HIV epidemic indicate a decrease in the incidence of the disease in the undiagnosed subpopulation over the past 10 years. A lack of access to care, however, has not been considered when modeling the population. Populations at high risk for contracting HIV are twice as likely to lack access to reliable medical care. In this thesis, we consider three contributors to the HIV population dynamics: susceptible pool exhaustion, lack of access to care, and usage of anti-retroviral therapy (ART) by diagnosed individuals. An extant problem in the mathematical study of this system is deriving parameter estimates due to a portion of the population being unobserved. We approach this problem by looking at the proportional change in the infected subpopulations. We obtain estimates for the proportional change of the infected subpopulations using hierarchical Bayesian statistics. The estimated proportional change is used to derive epidemic parameter estimates for a system of stochastic differential equations (SDEs). Model fit is quantified to determine the best parametric explanation for the observed dynamics in the infected subpopulations. Parameter estimates derived using these methods provide interpretability and recovery of the system. Simulations suggest that the undiagnosed population may be larger than currently estimated without significantly affecting the population dynamics.

Chapter 1. Bayesian Statistics and Mathematical Biology

1.1. Mathematical modeling in biology

Advances in technology allow scientists to collect and analyze massive data sets. Utilizing these data sets requires training in fields related to computer science, mathematics, and statistics. The field demands additional computational skills out of biologists, trending away from the traditional paradigm of primarily experimental work. These trends create a subset of modelers that are not quite traditional applied mathematicians, aren't statisticians, and are mostly self-taught programmers. An existing challenge is providing this growing population with the myriad of techniques available outside the training methods available to them. The NSF's Directorate for Biological Sciences (BIO) found that the unmet needs of its funded investigators centered around computational and mathematical training [2]. The NSF recently began a 'Big Ideas' program to address problems that require long-term solutions, and not surprisingly this issue is in focus [1].

One area in particular where these resources have been lacking is Bayesian statistics. A solution to this problem has been developing computational tools that are easy to use [5, 6]. These tools perform Bayesian inference on SMBL (systems biology markup language) models. SMBL models are procedurally generated, converting labeled wire diagrams of biological systems into models that can be simulated.

One issue facing widespread usage of these kinds of tools is that to seek them out, you need to have some idea of the problems existing with the tools you currently use. Current parameter estimation methods get the job done, and modelers may not be motivated to look elsewhere, especially considering the hurdles to utilizing these methods - additional programming, more computationally intensive, and a lack of background knowledge. However, obtaining estimates of parameter distributions instead of point estimates, or capturing the variability of a system or stochasticity of a process motivate increasing the usage of

Bayesian statistical inference in mathematical biology.

1.1.1. Approaches to modeling biological systems

Mathematical biology as a field attempts to represent biological systems directly with equations, while managing to simplify these systems to allow for mathematical or numerical analysis. The observed data D is a function of the response Y over time t. Normally Y consists of a set of n observable and unobservable components, and the length of the time under observation t is imposed based on financial or temporal constraints, or prior beliefs about the system behavior. Mathematical biologists try to construct Y so that it directly represents the system in question. Two famous models that illustrate this is the Lotka-Volterra model of population growth and the Michaelis-Menten model of enzyme kinetics.

- **Population modeling**

The Lotka-Volterra model represents a system of predator and prey species. The objective of the model is to understand how predator and prey dynamics affect each other, and this objective is used to simplify the model. We consider some birth rate of the prey population α, and some death rate β as a result of the predator population. The natural death rate of the prey is ignored for simplification purposes, as we assume that α is much larger. Finally, we assume that the predator population can only grow at some rate $\delta \leq \beta$, if prey are available as their primary or sole food source, and the predator die at some rate ϵ. This provides the following equations:

$$\frac{dx}{dt} = \alpha x - \beta xy$$

$$\frac{dy}{dt} = \delta xy - \epsilon y$$

This model contains many simplifications of a very complex, large system with variability, but it allows for biological inference on data of few observables. Most of the parameters are not observable, but we can roughly count the number of prey or predators at a given time to estimate the value of these rate constants; or predict the ability of a population to survive predation when assuming some parameter values.

- **Kinetics modeling**

  The Michaelis-Menten model of enzyme kinetics represents the following system:

$$E + S \rightleftharpoons ES \rightarrow P + E$$

This system consist of an enzyme E combining with a substrate molecule S. These molecules bind and separate at some rate k_{on} and k_{off}. When they exist in the ES conformation, the product molecule P is formed at some rate k_{cat}, and we assume that the enzyme separates and is again able to bind to a substrate. We also assume that all these rates are constants, and that we are looking at homogeneous populations of E, S, and P. In reality, some E, S, P will have different parameters describing their behavior, and these features are time and space dependent. A model based on this system is as follows.

$$\begin{aligned}
\frac{dE}{dt} &= -k_{on}E \cdot S + k_{off}ES + k_{cat}ES \\
\frac{dS}{dt} &= -k_{on}E \cdot S + k_{off}ES \\
\frac{dES}{dt} &= k_{on}E \cdot S - k_{off}ES - k_{cat}ES \\
\frac{dP}{dt} &= k_{cat}ES
\end{aligned}$$

The major contribution of Michaelis-Menten was not this system of equations, but rather another simplification. They showed that for the majority of enzyme systems, it appeared you could ignore the change in the ES population. This is because in general the

amount of ES reaches some level pretty quickly, and if the number of S molecules present is much larger than the number of E molecules, the change is not appreciable. Setting $\frac{dES}{dt} = 0$, the following model can be obtained:

$$\nu = \frac{dP}{dt} = \frac{V_{max}S}{K_M + S}$$

The value V_{max} represents the maximum velocity observed in the production of P, and K_M refers to the amount of S required to achieve half of the maximum velocity.

- **Different approaches to handle parameter problems**

In both of these models, we simplify the system in an attempt to obtain a manageable equation that contains the observables. In the Lotka-Volterra model, we minimize the number of unknown parameters since we can only get a rough estimate of the population of predator or prey. In the Michaelis-Menten model, we formulate our equation to be time-independent, containing two observables (V_{max} and S) and one unknown parameter.

Historically, clever simplifications and reductions have been applied to systems in order to make them tractable. Parameter estimation and identifiability in models with large numbers of parameters is challenging and computationally intensive. Biological networks can have disparate issues when it comes to available data. Some biological networks are very difficult to observe, such as large populations; some networks are not observable, such as what happens inside of a cell. These cases require special treatment to consider how to estimate the parameters, and whether we can estimate them at all. Other biological systems have large data sets associated with them. Bayesian statistical methods can be used to address many of these existing issues.

1.2. Bayesian statistics and parameter estimation

Bayes' rule describes the probability of event A occurring given event B occurred.

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

The probability of A occurring given B occurred is equivalent to the probability of event B occurring given A is occurred, multiplied by the probability of event A independent of event B, divided by the probability of event B occurring independent of event A. In Bayesian statistics, we take $\Pr(A)$ as the initial belief in the probability of A occurring. Parameter estimation for biological models also considers the current evidence and *a priori* beliefs about the system when constructing equations and considering parameters.

1.2.1. Handling big data

It has become relatively common place to collect data on the effect of experimental conditions on the expression of the entire genome within a large population of cells. It is also possible to use fluorescent tagging of proteins or cellular components to track the behavior of a large population of cells visually. One example of a Bayesian approach to handling this type of data is the Naive Bayes classifier [10]. The classifier was used to determine the log odds of different cellular markers in determining if a cell was likely to die or to continue reproducing. The methodology allowed the researchers to avoid making assumptions about the connections within this biological network, instead using Bayesian statistics to determine the likely contributors to the cell's "decision" to reproduce or to die.

1.2.2. Parameter estimation for large biological networks

Traditional parameter estimation for differential equation models consists of analytically solving equations for some initial conditions or knowns. This is not feasible for larger nonlinear models. In this case various computational techniques are often used, such as

least-squares fitting or gradient descent. These methods involve setting initial parameter estimates and parameter bounds, ideally based on previous experimental data.

A problem with this approach is that we are often more confident in some parameters than in others. Bayesian parameter estimation techniques are an improvement on such existing techniques by allowing the incorporation of parameter uncertainty [12].

The goal of biological mathematical modeling is normally not just producing accurate simulations, but also generating new hypotheses. The proteins JAK2 and STAT5 form a feedback loop that prevents treatment of some breast cancers [11]. A mathematical model of this system has 29 components, 25 equations, and 113 parameters [9]. In total, 513 data points were available to analyze, obtained from 24 different experimental conditions, making it a challenging problem to obtain parameter estimates. The authors used Markov Chain Monte Carlo to obtain Bayesian inference on the parameters. The model then allowed them to identify new observables in the system, generating new hypotheses.

1.2.3. Studying stochastic systems

Many biological systems may be better represented as stochastic systems, from the biochemical level to the population level [7, 8]. Some non-Bayesian methods available to study these systems include various stochastic simulation algorithms, like the Gillespie algorithm. However, it is challenging to estimate parameters for stochastic models without Bayesian statistics [12].

Bayesian methods for stochastic parameter estimates include using the stochastic nature of the model to create a likelihood, to allow for Bayesian inference on these models. Approximate Bayesian computation (ABC) uses the difference between the data and stochastic simulations of the model to replace the likelihood. In the following chapter, we

use Bayesian statistics to estimate two parameters that summarize the available information on two HIV populations. Then we construct several stochastic differential equation models and numerically solve to obtain parameter estimates for these equations. We compare our stochastic simulations of these models against the data, and quantify the goodness-of-fit by calculating the likelihood of observing the data given the model.

Chapter 2. Estimating Epidemiological Parameters of a Stochastic Differential Model of HIV Dynamics Using a Hierarchical Bayesian Model

2.1. Introduction

The human immunodeficiency virus (HIV) progresses in three stages. The first stage lasts approximately 3 months and individuals in this stage are approximately 10 to 25 times more effective at transmitting the disease [29]. The chronic stage can last from 5-10 years without medication [38].This is followed by acquired immunodeficiency syndrome (AIDS) and death shortly thereafter [29]. Individuals with HIV may go many years without diagnosis, during which time they may expose uninfected individuals to HIV. Efforts to improve the diagnosis rate include educational programs, as an individual's perceived risk was shown to be highly correlated with the individual obtaining multiple HIV tests [19,22,24,35]. Several studies have found a 50% reduction in risky behaviors after diagnosis, including safer sex practices, reduction in partner number, and medications that reduce viral load [27,39]. Diagnosis events resulting in behavior modification are not thought to be sufficient to eradicate the epidemic [20, 27].

After diagnosis, infected individuals have the opportunity to take anti-retroviral therapy (ART) that reduces their viral load and retards the progression of the disease. The earlier that ART is received the higher the reduction in transmission events. ART therapies could eradicate the epidemic in a population with high prevalence of infection even without the additional effect of behavioral changes [20]. Mathematical models estimate that the HIV epidemic could be reduced to less than 1% of the population infected (elimination phase) with universal testing and by providing ART consistently to newly diagnosed individuals [36]. However, issues with adherence and resistance are well documented in the

This chapter was originally published as Renee Dale and BeiBei Guo "Estimating epidemiological parameters of a stochastic differential model of HIV dynamics using hierarchical Bayesian statistics." *PLOS ONE* 2018 13(7): e0200126. Published by PLOS.

literature [13–15, 25, 28, 32]. Patients tend to report their adherence as much higher than it actually is, but studies indicate that even low adherence may be sufficient for control of the epidemic [13, 39]. Transmission is rare for individuals on ART, even with relatively high plasma HIV concentrations [30, 33].

The largest barrier to eradication of the epidemic is lack of access to care, including diagnostic services and ART costs or prescriptions. A lack in access to care could create pockets of undiagnosed individuals while the overall trend appears to be a reduction in the size of the epidemic [16]. Various studies report between 50 - 96% of diagnosed individuals in the U.S. rely on public medical care to obtain their ART medications [18, 21, 32, 34, 39]. Access to care remains critical, but this has not been considered when modeling the dynamics of the epidemic [19, 22].

Estimates using CD4 levels of newly diagnosed individuals suggest that the undiagnosed population is decreasing between 2005 - 2013 [37, 38]. CD4 levels can be used to estimate the progression of HIV [40]. We consider three possible causes for this decrease including exhaustion of the susceptible population. The size of the susceptible population is not easy to estimate since it depends on behavior risk. High risk populations include individuals in poverty and men who have sex with men (MSM) [38]. This is particularly critical in the southern U.S., where individuals tend to be poor and lack access to medical care [16,38]. As the at-risk population decreases the number of diagnoses will also decrease, which will cause the estimated number of undiagnosed individuals to decrease.

An additional possibility is that the reduction in number of diagnoses is due to individuals lacking access to care. HIV is over-represented in impoverished populations where access to diagnosis and treatment may be more difficult to obtain. In this case, the number of newly diagnosed individuals is not representative of the number of undiagnosed individ-

uals, and the estimates will be inaccurate. Finally, the usage of anti-retroviral therapies reduces the viral load and transmission potential of infected individuals.

The difficulty in studying this system mathematically lies in parameter estimates. A minimal model of this system requires at least three parameters: transmission of the disease, diagnosis of the disease, and death due to the disease. Since knowledge about the undiagnosed population is restricted to those who have been diagnosed, estimates of these parameters are generally forced to assume that this population is representative of the whole. Our motivation to model the system stochastically arises from heterogeneity due to reporting delays associated with population-level data [31]. Stochastic modeling will allow us to better understand both the effectiveness of our estimates and the quality of model fit.

In this work we use coupled statistical and mathematical methodology to study the relationships between the three hypothesized causes and their respective population dynamics. We use hierarchical Bayesian statistics to get estimates for the size of the infected populations and their proportional changes across the years. These estimates are used to calculate epidemiological parameters for a system of stochastic differential equations. Currently we are not aware of any similar methods in the literature. Such a problem is challenging as the proportional change across the populations is a hyperparameter controlling the yearly proportions, which each have their own statistical model. This results in a large model that must be studied numerically.

The resulting simulations give insight into the implications of the estimated undiagnosed population on epidemiological parameters. Our model suggests that the undiagnosed population may be larger than current estimates while recovering population dynamics. The best recovery occurs when the increase in the diagnosed population due to diagnosis

is greater than the decrease in the undiagnosed population. We hope this study will help inform future efforts to improve the situation of infected individuals and prevent future outbreaks.

2.2. Materials and methods

2.2.1. Bayesian statistics

A Markov model where p_t centered at qp_{t-1} is used to estimate the proportional change in the infected populations over time, where p_t is the proportion in the current year and q is the proportional change. These random variables are estimated using Bayesian statistics.

The sampling model is $x_t \sim Bin(n_t, p_t)$, where n_t is population size in the current year. The random variable q is taken as a hyperparameter for p_t. The random variables to be estimated for each infected subpopulation are q and p_t, where t = 2005, ..., 2013. We estimate the random variables of undiagnosed and diagnosed subpopulations independently.

- **Prior**

The prior for the proportional change q is a gamma distribution.

$$\pi(q) \propto q^{\alpha-1}e^{-\beta q}$$

The parameters were chosen so that the prior distribution was centered at the arithmetic estimates of q obtained from the CDC [38]. The arithmetic estimates were obtained by calculating:

$$\frac{1}{n}\Sigma_{i=2}^{n}\frac{p_i}{p_{i-1}}$$

The arithmetic estimate for the undiagnosed q (q_u) is 0.979, and for the diagnosed q (q_d) 1.025. The priors used were GAM(9.79,10) and GAM(10.25,10) so they were centered at 1.

The prior for the random variable p_t, the undiagnosed proportion, is a beta distribution centered at the previous proportion times q. The parameters of the beta distribution are $\alpha = 0.1n_{t-1} \times qp_{t-1}$, and $\beta = 0.1n_{t-1} - \alpha$.

$$\pi(p_t) \propto p_t^{\alpha-1}(1-p_t)^{\beta-1}$$

In the case where $t = 1$, the previous undiagnosed proportion is taken to be the expert opinion of 20%, and the diagnosed proportion to be $1-p_0(\textit{undiagnosed})$ [23]. The prior for the diagnosed population is formulated in the same way. Population sizes were considered in units of thousands.

- **Likelihood**

The likelihood is a binomial distribution, representing the chance of selecting an undiagnosed or diagnosed individual at random from the total infected population. For a given year t, the proportion of undiagnosed individuals depends on the total number of individuals:

$$\mathcal{L}(p_t|x_t, n_t) \propto p_t^{x_t}(1-p_t)^{n_t-x_t}$$

where x_t is the total number of undiagnosed or diagnosed individuals, and n_t is the total number of infected individuals. The likelihood across all the years is the product of each year's likelihood.

$$\mathcal{L}(p_1, p_2, ..., p_9|x_1, x_2, ..., x_9, n_1, n_2, ..., n_9) \propto \Pi_{t=1}^{9} p_t^{x_t}(1-p_t)^{n_t-x_t}$$

The likelihood for the diagnosed population was formulated in the same way.

- **Posterior**

The joint posterior distribution is proportional to the priors multiplied by the likelihoods for all 9 years:

$$f(p_1, p_2, ..., p_9, q) \propto \pi(q) \times \Pi_{t=1}^{9} \pi(p_t|q, p_{t-1}) \times \mathcal{L}(p_1|x_{t=1}) \times \mathcal{L}(p_2|x_{t=2}) \times ... \times \mathcal{L}(p_9|x_{t=9})$$
$$\propto q^{\alpha-1} e^{-\beta q} \times \Pi_{t=1}^{9} \Big(p_t^{\alpha-1} (1-p_t)^{\beta-1} \times p_t^{x_t} (1-p_t)^{n_t - x_t} \Big)$$

The posterior full conditional of p_t for $t = 2005, ..., 2012$ is:

$$f(p_t|q, p_{t-1}, p_{t+1}) \propto \Pi_{t=1}^{9} \Big(\mathcal{L}(p_t|x_t) \times \pi(p_t|q, p_{t-1}) \times \pi(p_{t+1}|q, p_t) \Big)$$

The posterior full conditional of 2013, the 9th year, is:

$$f(p_9|x_9, p_8, q) \propto \mathcal{L}(p_9|x_{t=9}) \times \pi(p_9|q, p_8)$$

The full conditional of q does not have a closed form. The forms of the diagnosed random variables are the same. Random variable estimates were obtained using Metropolis-Hastings nested within a Gibbs sampler over 100,000 iterations with R version 3.3.3 [41]. The proposal distribution was a truncated normal distribution, using package rmutil [42], centered at the previous value of the parameter. Proportions 1 through 9 were sampled consecutively, followed by hyperparameter q. The trace plots converged quickly, and the first 2000 samples were removed. Code is provided in Appendices B and C.

2.2.2. Stochastic differential equations

The hyperparameter q was estimated to be 0.978 for the undiagnosed population and 1.036 for the diagnosed population. These were used as a boundary to solve for the epi-

demiological parameters in a simple stochastic differential model.

$$dU = (q_u - 1)Udt + d\omega_t dt$$

$$dD = (q_d - 1)Ddt + d\omega_t dt$$

where U is the undiagnosed and D is the diagnosed populations, and $d\omega_t \sim Nor(0, \sigma)$ is Brownian white noise with units $\sqrt{(t)}$. The variance σ is chosen to be 10% of the size of the population.

The simplest model is constructed describing the dynamics of the infected subpopulations. The values of parameters transmission (τ), diagnosis (δ), and death (ϵ) are calculated using the constraining q:

$$\begin{aligned} dU =& \big[\tau(U + D) - \delta U - \epsilon\big]dt && (2.1)\\ \cong& (1 - q_u)Udt + d\omega_t dt && (2.2)\\ =& -0.022Udt + d\omega_t dt && (2.3)\\ dD =& (\delta U - \epsilon D)dt && (2.4)\\ \cong& (1 - q_d)Ddt + d\omega_t dt && (2.5)\\ =& 0.036Ddt + d\omega_t dt && (2.6) \end{aligned}$$

We consider the parameters pseudo-steady state, and use the 2005 population sizes to estimate them. In addition, we assume that the general population are at steady state, and consider only the increased death rate due to infection ϵ as 0.01 [38]. The diagnosis rate δ is estimated by:

$$\delta U = q_d D + \epsilon D = \frac{0.046 D_{2005}}{U_{2005}} = 0.165$$

and the transmission rate τ is estimated by:

$$\tau(U + D) = (1 - q_u)U + \delta U + \epsilon U$$

$$\tau = \frac{0.153U_{2005}}{U_{2005} + D_{2005}} = 0.0334$$

Due to the magnitude of the scale of this system we assume that all events will happen, and the source of the stochasticity is primarily reporting issues. Tau leap algorithm was used to preform the stochastic simulations. A time step of 1 year was selected, and the population at time *t+1* is the numerical solution of the population at time t and random noise from a $NOR(0,\sigma)$, where σ is 10% of the population at time t_0 with units $\sqrt{t}$. The initial conditions for the infected populations were sampled from the posterior distributions obtained by the Bayesian estimates. Calculations were performed in Matlab [43] and code is available upon request.

The diagnosis rate was calculated using data from [37]. The susceptible population is estimated as twice the national average rate of self-identified MSM among adults. The mortality rate increase due to HIV was estimated using data from [38]. All calculations, including the effective parameter rates, are provided in Appendix A (in progress).

2.3. Results

2.3.1. Bayesian model

Bayesian estimates for the proportions of diagnosed or undiagnosed individuals was obtained concurrently with the estimated proportional change. The prior distribution was chosen to be a beta for the proportions and a gamma for the proportional changes. The likelihood function was a binomial, representing the chance of randomly selecting a diagnosed or undiagnosed individual from a pool of infected individuals. The posterior did not have a closed form. Due to the symmetry of the posterior samples we summarize them using their mean and variance. The posterior means of the proportions for both undiagnosed

and diagnosed estimates are very close to the original data (Fig. 2.1) [37]. The posterior mean of q_u is 0.96, and q_d is 1.02. This means that 96% of the undiagnosed population is preserved from year to year, or is dropping by about 4% per year. Similarly, the diagnosed population is increasing by 2% per year. Posterior means and variances are given in Table 2.1.

Table 2.1. Summary statistics of the posterior distribution. The parameter p represents the estimated size of the proportion in that year. The hyperparameter q represents the estimated proportional change of the population across all years.

Undiagnosed			**Diagnosed**		
Parameters	**Mean**	**Variance**	**Parameters**	**Mean**	**Variance**
p_{2005}	0.78	0.0006	p_{2005}	0.22	0.0006
p_{2006}	0.79	0.0011	p_{2006}	0.21	0.0006
p_{2007}	0.80	0.0012	p_{2007}	0.20	0.0006
p_{2008}	0.81	0.0012	p_{2008}	0.16	0.0006
p_{2009}	0.82	0.0012	p_{2009}	0.19	0.0006
p_{2010}	0.82	0.0012	p_{2010}	0.18	0.0006
p_{2011}	0.83	0.0012	p_{2011}	0.18	0.0006
p_{2012}	0.83	0.0012	p_{2012}	0.17	0.0005
p_{2013}	0.84	0.0012	p_{2013}	0.16	0.0005
q_d	1.036	0.1046	q_u	0.978	0.0978

2.3.2. Stochastic differential model

The Bayesian estimates of the proportional change in the diagnosed and undiagnosed population from 2005 to 2013 were used to determine the epidemiological parameters for a system of stochastic differential equations. The parameters transmission (τ), diagnosis (δ), and death (ϵ) were calculated using the proportional changes in the respective population.

$$dU = (q_u - 1)Udt + d\omega_t dt = \big[\tau(U + D) - \delta U - \epsilon U\big]dt \tag{2.7}$$

$$dD = (q_d - 1)Ddt + d\omega_t dt = \big[\delta U - \epsilon D\big]dt \tag{2.8}$$

where U is the undiagnosed and D is the diagnosed populations, and $d\omega_t \sim N(0, \sigma)$ is the noise term. These base estimates fit the data very well (Fig. 2.2).

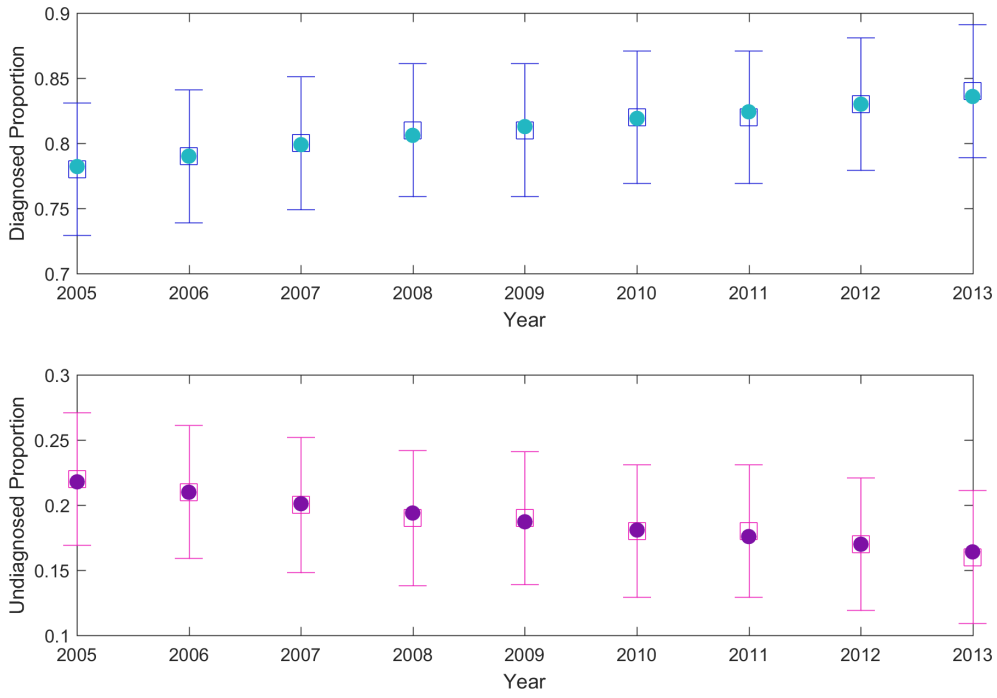

Figure 2.1. Posterior information obtained from hierarchical Bayesian statistics. Bayesian estimates are shown as hollow squares with error bars showing standard deviations. Estimated proportion of diagnosed (pink) and undiagnosed (blue) populations recover the estimated proportions (circles). Data obtained from [37].

- **Exhaustion of susceptibles**

In the case where the susceptible population is not much larger than the infected population, the transmission is dependent on the size of both populations. We estimate the susceptible population size as a fraction of the total infected population:

$$S = fT$$

Then this is substituted into the model. The transmission term becomes

$$\tau TS \cong \tau f T^2$$

This gives an effective increase of $f\tau$ in the transmission rate - see Table 2.2. This increase causes the simulations to fail to recover the diagnosed and undiagnosed popula-

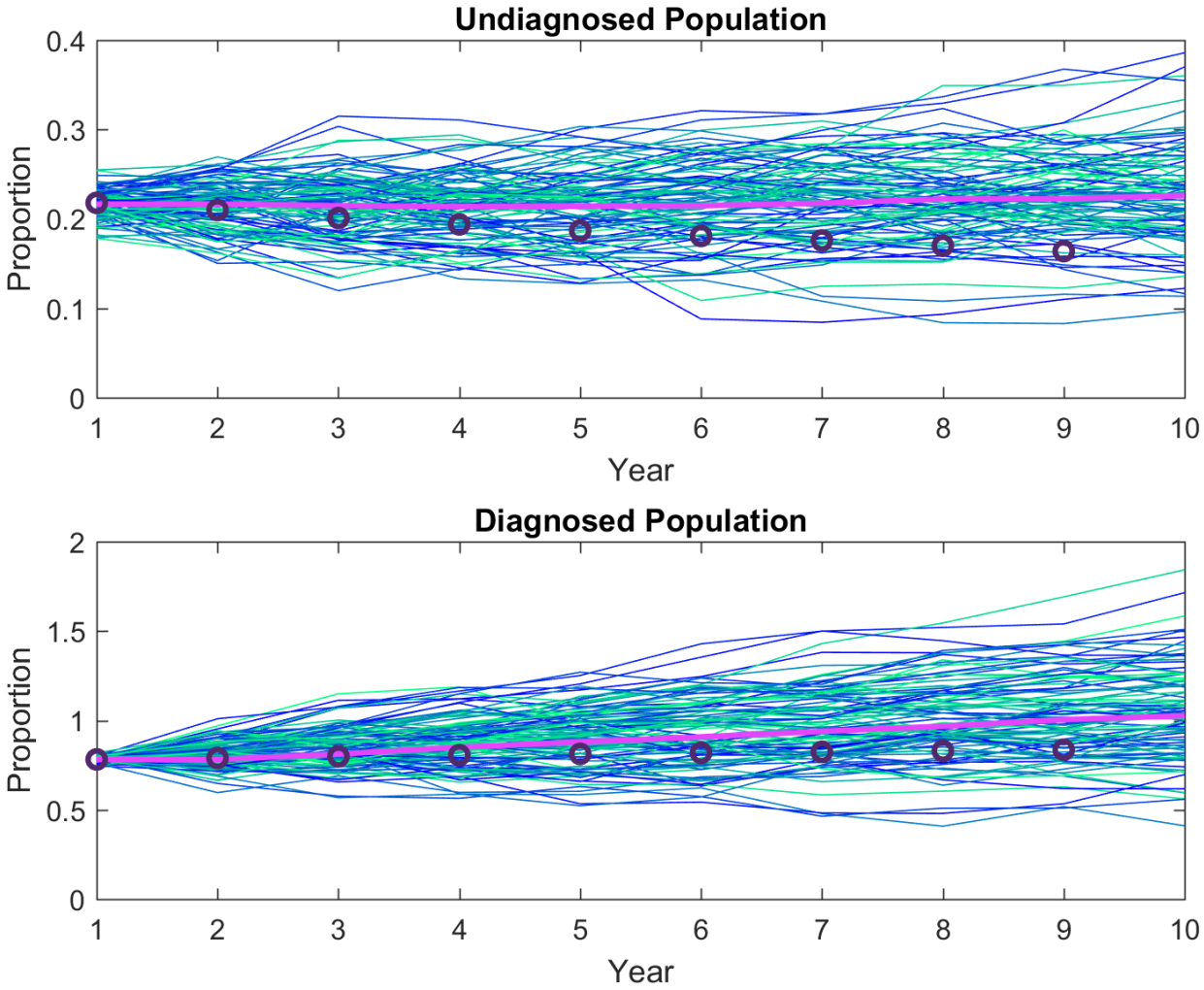

Figure 2.2. Method Validation. The method was tested by simulating with the epidemiological parameters calculated using the Bayesian estimates of the proportional changes as constraints. The mean of 100 stochastic simulations (pink line) is compared with the data (circles). Proportions are relative to initial proportion.

tion dynamics, although the susceptible population does decrease significantly (Fig. 2.3). This result is intuitive since the infection rate is increased, but the diagnosis rate is not representative of this rate.

- **Lack of access to care**

Lack of access to care may be conceptualized as pockets of undiagnosed individuals who are not being diagnosed. To capture this, we consider the diagnosis rate to be independent of the size of the undiagnosed population. The diagnosis rate is estimated as:

$$\delta U = q_d D + \epsilon D = \frac{0.046 D_{2005}}{U_{2005}} \cdot U_{2005} = 0.036$$

The resulting equation for the undiagnosed subpopulation then becomes:

$$dU = \big[\tau(U + D) - \delta_0 - \epsilon U\big] dt$$

where δ_0 is 0.036. This large reduction in the diagnosis rate recovers the population dynamics well (Fig. 2.4).

- **ART usage**

Since ART results in a viral load that has low chance of infecting a susceptible individual, we removed these individuals from the pool of infected individuals able to transmit the disease. Since 96% of diagnosed individuals reported taking anti-retroviral therapies in a previous study, the transmission term was modified as follows [39].

$$\tau\big[U + (1 - 0.96)D\big]$$

Variable or poor adherence on the part of diagnosed individuals is ignored due to the body of literature indicating that large benefit is gained from even poor adherence [13, 25].

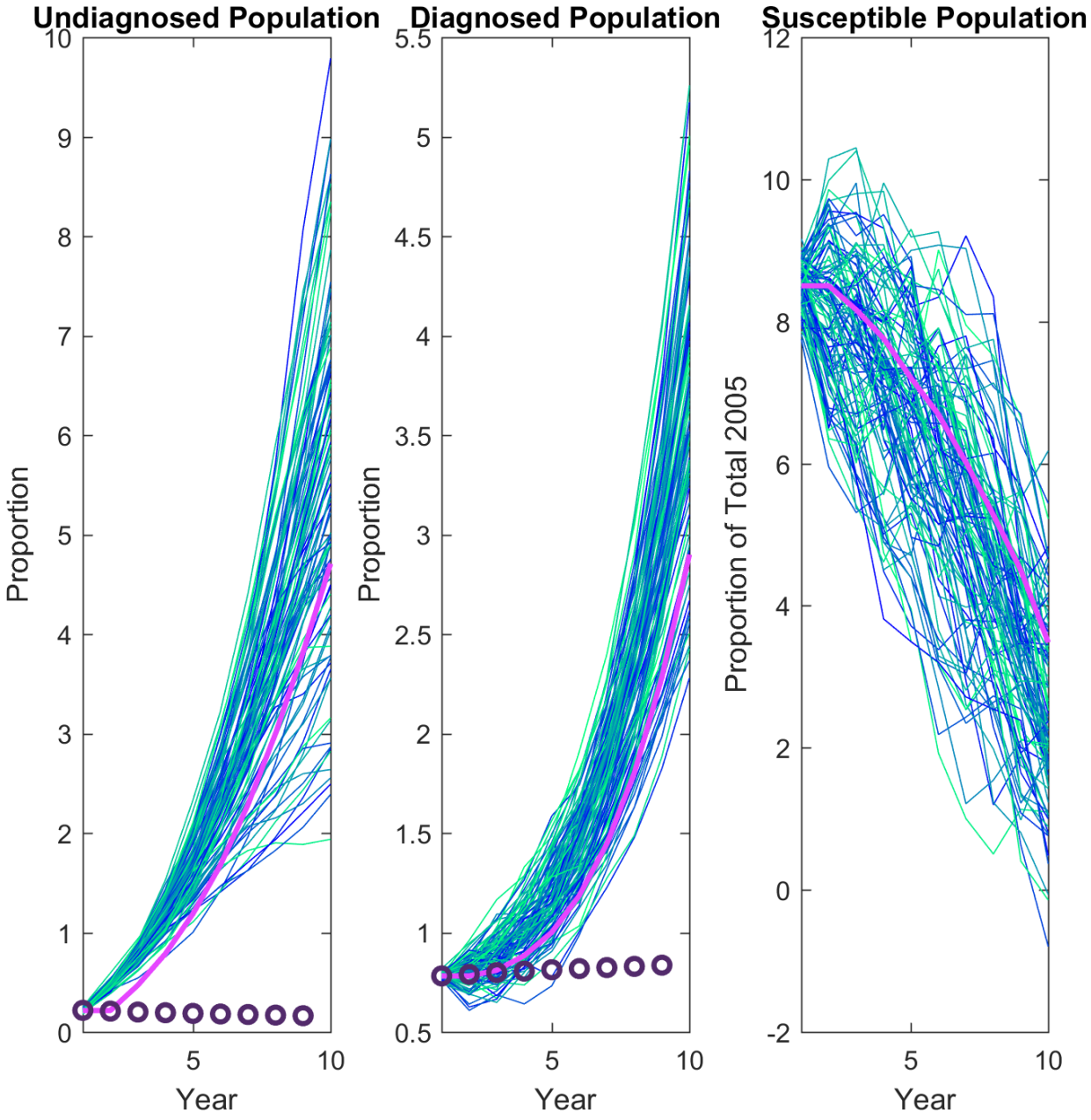

Figure 2.3. Exhaustion of Susceptibles. Transmission of the disease is altered to reflect the impact of the size of the susceptible population. The mean of 100 stochastic simulations (pink line) is compared to the data (circles). Proportions are relative to initial proportion.

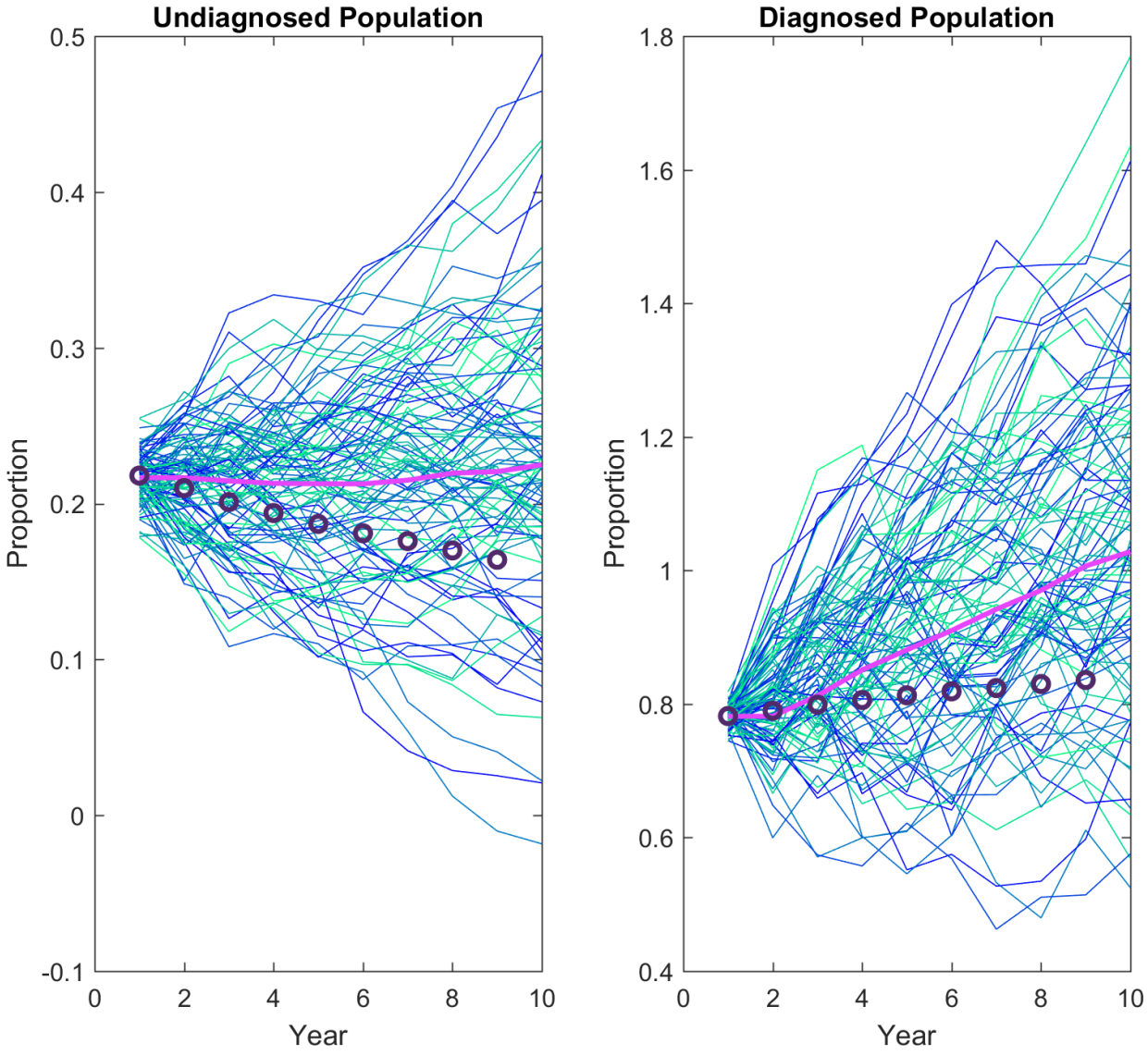

Figure 2.4. Lack of Access to Care. The effect of undiagnosed individuals lacking access to care affects the rate of diagnosis of the undiagnosed individuals. The diagnosis rate is held constant to reflect this scenario. The mean of 100 stochastic simulations (pink line) is compared with the data (circles). Proportions are relative to initial proportion.

This gives good recovery of both subpopulation dynamics and agrees best with both undiagnosed and diagnosed estimates (Fig. 2.5).

2.3.3. Multiple causes

Since it seems likely that most or all of these scenarios affect the infected population simultaneously, we analyze all their possible combinations (Appendix A). The parameters were altered as described in Table 2.2. To determine the best cause, we quantify the goodness of fit by determining the relative likelihood of observing the data given the mean and variance of the stochastic simulations. These probabilities are given in Fig. 2.6 with numerical details in the Appendix A table, as well as the average probability over the 9 years.

Table 2.2. Transmission and diagnosis rates are different under the different hypotheses. Average likelihood across both populations and all years (Fig. 2.6, Supplemental Table in Appendix A.

Model	**Transmission Rate**	**Diagnosis Rate**	**Likelihood**
Base model	$\tau(U+D)$	δU	0.83
Exhaustion of Susceptibles (ES)	$\tau f(U+D)^2$	δU	0.26
Lack of Access to Care (LAC)	$\tau(U+D)$	δ_0	0.88
Anti-retroviral Therapies (ART)	$\tau(U+0.04D)$	δU	0.75
ES and LAC	$\tau f(U+D)^2$	δ_0	0.54
ES and ART	$\tau f(U+0.04D)^2$	δU	0.52
LAC and ART	$\tau(U+0.04D)$	δ_0	0.78
ES, LAC, and ART	$\tau f(U+0.04D)^2$	δ_0	0.58

Lack of access to care, ART usage, or their combined resulted in the best recovery of the data for both the undiagnosed and diagnosed populations. Under the ART scenario, the diagnosed population has been reduced by 96%, resulting in a dynamic reduction in the transmission rate. We originally estimate the transmission rate to be 3.3% of the total infected population. This is close to the literature estimate of around 4% [17, 23]. With the majority of the diagnosed population removed, the effective transmission rate is much lower. Under the LAC scenario, there is a constant diagnosis rate. This represents a yearly reduction in the undiagnosed population and increase in the diagnosed population of 3.6%.

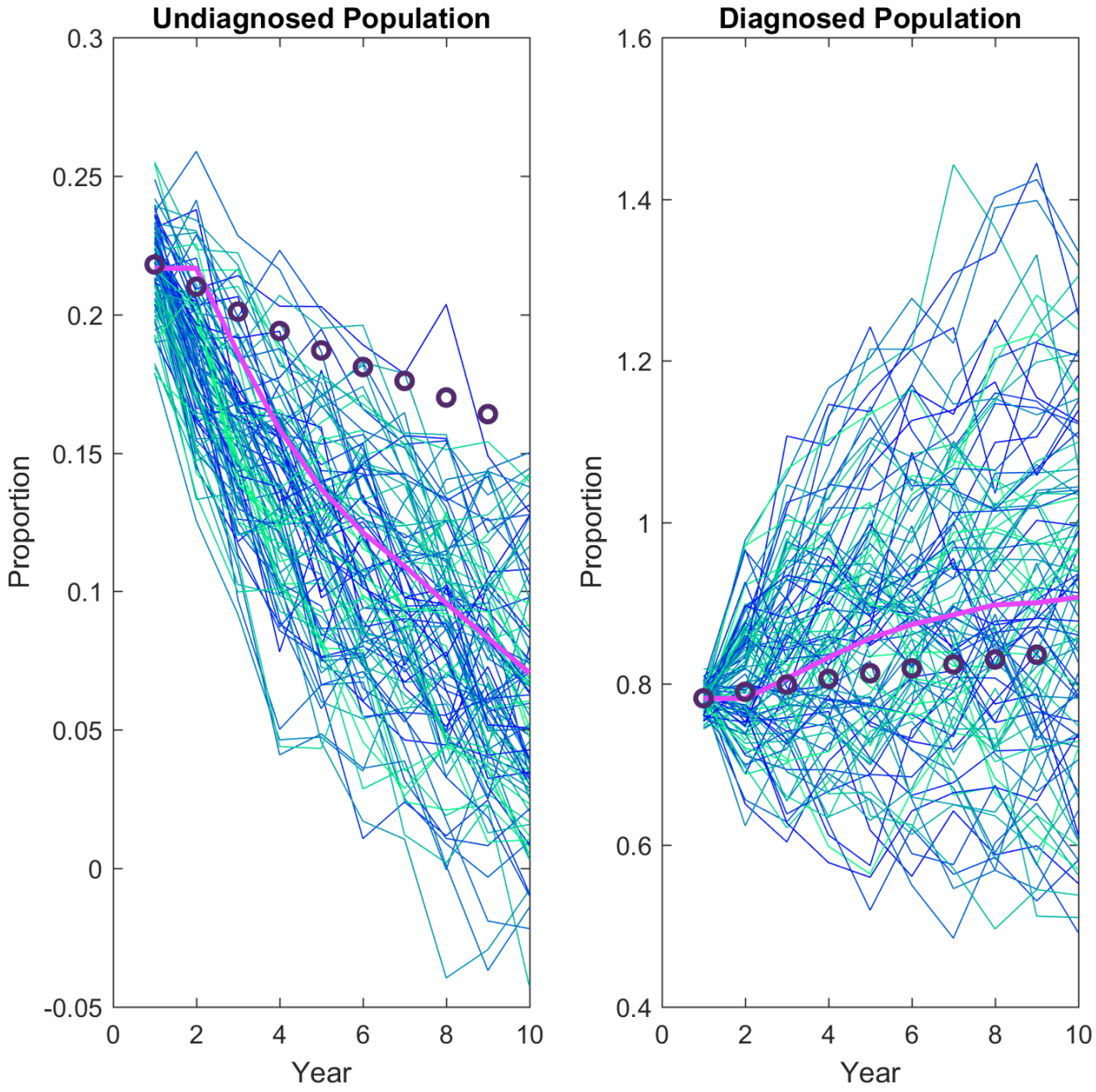

Figure 2.5. Anti-retroviral Therapy Usage. To reflect the high levels of ART prescription and usage reported by diagnosed individuals this percentage is removed from the pool of diagnosed individuals able to transmit the disease. The mean of 100 stochastic simulations (pink line) is compared to the data (circles). Proportions are relative to initial proportion.

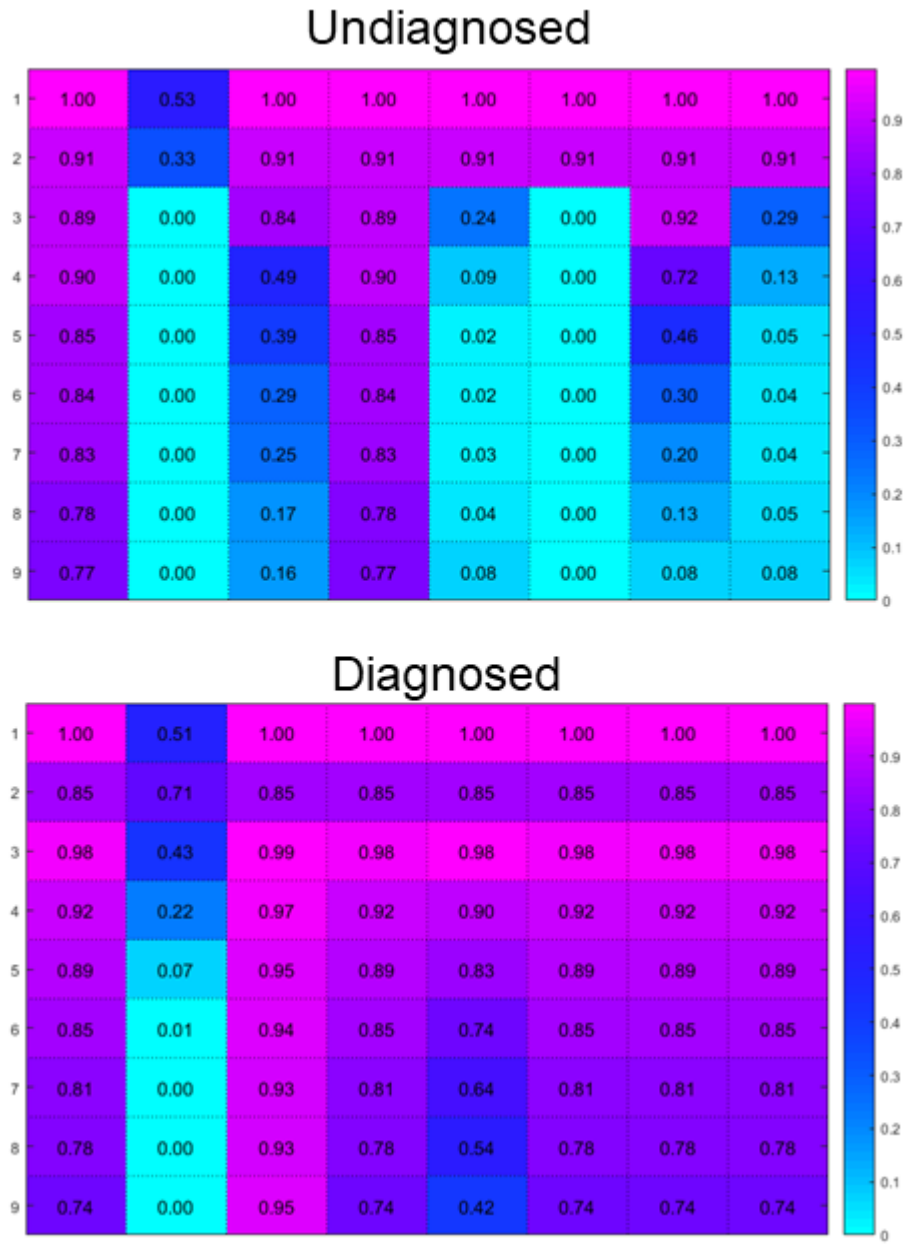

Figure 2.6. Model fit was quantified by calculating the relative likelihood of observing the data within the simulations. A higher likelihood is represented by a hotter color. From left to right: Base model, Exhaustion of Susceptibles (ES), Lack of Access to Care (LAC), Anti-Retroviral Therapy usage (ART), ES and LAC, ES and ART, LAC and ART, and ES, LAC, and ART. Details provided in Appendix A.

This means the diagnosed population grows faster than the undiagnosed population is reduced for a diagnosis event. Although lack of access to care in the undiagnosed population would mean the data are inaccurate, in our case the best fitting model is consistent for both subpopulations.

2.4. Discussion

We were able to obtain conservative estimates of the proportional changes in the diagnosed and undiagnosed HIV-infected populations using hierarchical Bayesian statistics. Our estimates suggest that the proportion of infected individuals who are undiagnosed is decreasing by approximately 2.2% each year from 2005 to 2013, while the proportion of diagnosed individuals is increasing by approximately 3.6%. We used the proportional change as constraints on a system of stochastic differential equations. This allowed us to estimate the transmission and diagnosis rates. We were able to recover reasonable parameter estimates and population dynamics using this methodology. To learn more about the cause of the decrease in the undiagnosed population, we considered some scenarios that would affect the epidemiological parameters: exhaustion of the susceptible population, lack of access to care, and reduction in viral load by anti-retroviral therapy.

We were able to recover the diagnosed population dynamics when we altered the parameters to reflect these scenarios with the exception of including exhaustion of susceptibles. Including the size of the susceptible population dramatically increased the transmission rate and caused the size of the infected populations to increase rapidly. In the other scenarios some interesting dynamics could be observed in the undiagnosed population. Lack of access to care was simulated by considering diagnosis rate a constant unaffected by the size of the undiagnosed population. This resulted in an improvement in the likelihood of observing the data (Fig. 2.6, Appendix A). Anti-retroviral therapy usage also improved the overall recovery, but this effect was weaker for the undiagnosed population dynamics. Although

the undiagnosed population size is dependent on the quality of the data available on the diagnosed population of that year, these results indicate that the scenarios that maximizes the probability of observing the diagnosed population also maximizes the probability of observing the diagnosed population estimates.

The observed results suggest that lack of access to care and ART usage contribute to the infected population dynamics. This is not unexpected. Many individuals with HIV are reported to lack access to care [16, 26]. In areas with high poverty rates the death rate of infected individuals is much higher than that of the general population [38, 44]. In 2017 the New York Times reported groups of untreated individuals in the deep south dying due to their lack of access to care [16]. The effect of simulating a lack of access to care suggest this to be a significant contributing factor to the infected population dynamics. Both models and studies have shown that providing ART to infected individuals in the early stages of HIV reduces transmission events and frequency of death due to AIDS [20, 28, 30, 32, 33, 36]. Even poor adherence may be enough to control or eradicate the epidemic and increase quality of life for infected individuals [14, 15, 25]. Greater effort must be made to ensure these populations have access to life-saving treatments.

Appendix A. Rights and Permissions for Previously Published Articles

Reuse of PLOS Article Content
Content Owned by Someone Else
Using Article Content Previously Published in Another Journal
Acceptable Licenses for Data Repositories
Removal of Content Used Without Clear Rights
Guidelines for Trademarks
Giving Proper Attribution for Use of Content
Give Feedback

Licenses and Copyright

The following policy applies to all PLOS journals, unless otherwise noted.

Reuse of PLOS Article Content

PLOS applies the Creative Commons Attribution (CC BY) license to articles and other works we publish. If you submit your paper for publication by PLOS, you agree to have the CC BY license applied to your work. Under this Open Access license, you as the author agree that anyone can reuse your article in whole or part for any purpose, for free, even for commercial purposes. Anyone may copy, distribute, or reuse the content as long as the author and original source are properly cited. This facilitates freedom in re-use and also ensures that PLOS content can be mined without barriers for the needs of research.

Content Owned by Someone Else

If your manuscript contains confidential information or content such as photos, images, figures, tables, audio files, videos, proprietary protocols, code, etc., that you or your co-authors do not own, we will require you to provide us with proof that the owner of that content (a) has given you written permission to use it, and (b) has approved of the publication of such information or content under the CC BY license. This form can be used to request permissions. Under no circumstances should your manuscript contain third party trade secret information.

If you do not have owner permission, we will ask you to remove that content and/or replace it with other content that you own or have such permission to use.

Don't assume that you can use any content you find on the Internet, or that the content is fair game just because it isn't clear who the owner is or what license applies. It's up to you to ascertain what rights you have—if any—to use that content.

Using Article Content Previously Published in Another Journal

Many authors assume that if they previously published a paper through another publisher, they own the rights to that content and they can freely use that content in their PLOS paper, but that's not necessarily the case – it depends on the license that covers the other paper. Some publishers allow free and unrestricted re-use of article content they own, such as under the CC BY license. Other publishers use licenses that allow re-use only if the same license is applied by the person or publisher re-using the content.

If the paper was published under a CC BY license or another license that allows free and unrestricted use, you may use the content in your PLOS paper provided that you give proper attribution, as explained above.

If the content was published under a more restrictive license, you must ascertain what rights you have under that license. At a minimum, review the license to make sure you can use the content. Contact that publisher if you have any questions about the license terms – PLOS staff cannot give you legal advice about your rights to use third-party content. If the license does not permit you to use the content in a paper that will be covered by an unrestricted license, you must obtain written permission from the publisher to use the content in your PLOS paper. Please do not include any content in your PLOS paper which you do not have rights to use, and always give proper attribution.

Acceptable Licenses for Data Repositories

If any relevant accompanying data is submitted to repositories with stated licensing policies, the policies should not be more restrictive than CC BY.

Removal of Content Used Without Clear Rights

PLOS reserves the right to remove any photos, captures, images, figures, tables, illustrations, audio and video files, or other confidential or proprietary content, from any article, whether before or after publication, if concerns are raised about copyright, license, or permissions and the authors are unable to provide documentation confirming that appropriate permissions were obtained for publication of the content in question under a CC BY license.

Guidelines for Trademarks

Ensure that any reference to a trademark (such as a brand name) is used as an adjective, and not a noun or verb. The trademark should be immediately followed by the generic term for the object that it modifies. Note that because a trademark cannot be used as a noun, it cannot be presented in the possessive or plural form. Please see the following example for reference:

Reuse of PLOS Article Content
Content Owned by Someone Else
Using Article Content Previously Published in Another Journal
Acceptable Licenses for Data Repositories
Removal of Content Used Without Clear Rights
Guidelines for Trademarks
Giving Proper Attribution for Use of Content
Give Feedback

Licenses and Copyright

The following policy applies to all PLOS journals, unless otherwise noted.

Reuse of PLOS Article Content

PLOS applies the Creative Commons Attribution (CC BY) license to articles and other works we publish. If you submit your paper for publication by PLOS, you agree to have the CC BY license applied to your work. Under this Open Access license, you as the author agree that anyone can reuse your article in whole or part for any purpose, for free, even for commercial purposes. Anyone may copy, distribute, or reuse the content as long as the author and original source are properly cited. This facilitates freedom in re-use and also ensures that PLOS content can be mined without barriers for the needs of research.

Content Owned by Someone Else

If your manuscript contains confidential information or content such as photos, images, figures, tables, audio files, videos, proprietary protocols, code, etc., that you or your co-authors do not own, we will require you to provide us with proof that the owner of that content (a) has given you written permission to use it, and (b) has approved of the publication of such information or content under the CC BY license. This form can be used to request permissions. Under no circumstances should your manuscript contain third party trade secret information.

If you do not have owner permission, we will ask you to remove that content and/or replace it with other content that you own or have such permission to use.

Don't assume that you can use any content you find on the Internet, or that the content is fair game just because it isn't clear who the owner is or what license applies. It's up to you to ascertain what rights you have—if any—to use that content.

Using Article Content Previously Published in Another Journal

Many authors assume that if they previously published a paper through another publisher, they own the rights to that content and they can freely use that content in their PLOS paper, but that's not necessarily the case – it depends on the license that covers the other paper. Some publishers allow free and unrestricted re-use of article content they own, such as under the CC BY license. Other publishers use licenses that allow re-use only if the same license is applied by the person or publisher re-using the content.

If the paper was published under a CC BY license or another license that allows free and unrestricted use, you may use the content in your PLOS paper provided that you give proper attribution, as explained above.

If the content was published under a more restrictive license, you must ascertain what rights you have under that license. At a minimum, review the license to make sure you can use the content. Contact that publisher if you have any questions about the license terms – PLOS staff cannot give you legal advice about your rights to use third-party content. If the license does not permit you to use the content in a paper that will be covered by an unrestricted license, you must obtain written permission from the publisher to use the content in your PLOS paper. Please do not include any content in your PLOS paper which you do not have rights to use, and always give proper attribution.

Acceptable Licenses for Data Repositories

If any relevant accompanying data is submitted to repositories with stated licensing policies, the policies should not be more restrictive than CC BY.

Removal of Content Used Without Clear Rights

PLOS reserves the right to remove any photos, captures, images, figures, tables, illustrations, audio and video files, or other confidential or proprietary content, from any article, whether before or after publication, if concerns are raised about copyright, license, or permissions and the authors are unable to provide documentation confirming that appropriate permissions were obtained for publication of the content in question under a CC BY license.

Guidelines for Trademarks

Ensure that any reference to a trademark (such as a brand name) is used as an adjective, and not a noun or verb. The trademark should be immediately followed by the generic term for the object that it modifies. Note that because a trademark cannot be used as a noun, it cannot be presented in the possessive or plural form. Please see the following example for reference:

Appendix B. Diagnosed Population R Code

```
# # # # # # #
library(MCMCpack)
library(pracma)
library(msm)
library(rmutil)
# # # # # #  DATA  # # # # # #
# Source: Song R, Hall HI, Green TA, Szwarcwald CL, Pantazis N. Using CD4 Data to
#Estimate HIV Incidence, Prevalence, and Percent of Undiagnosed Infections in
#the United States. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2017;74(1):3{9.
# # # # # # # # # # # #  # # #

total = c(896.9,923.2,951.5,979.7,1006.5,1032.6,1057.8,1082.1,1104.6)
diagnosed=c(701.3758,729.328,760.2485,789.6382,818.2845,845.6994,871.6272,898.143,923.44
x=diagnosed

# # # # # # # # # # # #  # # #

plot(total)
year=c(2005,2006,2007,2008,2009,2010,2011,2012,2013)
plot(year,undiagnosed,xlab=Year,ylab=,xaxt='n',yaxt='n')
axis(1,at=year,labels=year)
axis(2,at=undiagnosed,labels=undiagnosed)
title(main=Estimated Undiagnosed HIV Population)
plot(x)
v=matrix(0,8)
for (i in 1:8){v[i]=diagnosed[i+1]/diagnosed[i]}
mean(v)
var(v)

# # # # # #
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
# # # # # #

p0=.8 #expert prior
n0=.25*total[1]

posteriorp=function(pt,pt1,pt.1,nt,nt1,nt.1,xt,xt1,xt.1,q,f){
## this year t
talpha=pt.1*nt.1*f*q
```

```
tbeta=(1-pt.1*q)*nt.1*f

## next year t+1
t1alpha=pt*nt*q*f
t1beta=(1-pt*q)*nt*f
# use beta to approximate binomial to keep numbers unrounded
a=dbeta(pt,talpha,tbeta)*dbeta(pt,xt+1,nt-xt+1)
b=dbeta(pt1,t1alpha,t1beta)*dbeta(pt1,xt1+1,nt1-xt1+1)
a*b
}

posterior_individ=function(pt,pt.1,xt,xt.1,nt,nt.1,q,f){
talpha=pt.1*nt.1*f*q;  tbeta=(1-pt.1*q)*nt.1*f
dbeta(pt,xt+1,nt-xt+1)*dbeta(pt,talpha,tbeta)
}

posterior_q=function(pt,pt.1,xt,xt.1,nt,nt.1,q,f){
talpha=pt.1*nt.1*f*q;  tbeta=(1-pt.1*q)*nt.1*f
dbeta(pt,talpha,tbeta)
}

# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #
# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #
# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #

##### GIBBS SAMPLER #####
sample=100000
gibb.sample=matrix(0,sample,10) #columns = parameters
tmu=matrix(.9,9,1) # initial values for proportion
q=1 # initial value for q
var_star = 0 # dummy variable for each iter

# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #

sd.rw=matrix(.05,11);accept=matrix(0,11) # acceptance matrix
sd.rw[10]=.52
f=.1 # to scale the strength of the prior
for (i in 1:sample ){
set.seed(i) # for reproducibility
# # # # # # # # # # # # # # #
## p1 : marginal posterior involves p0 and p1
```

```
var_star <- rtnorm(1,tmu[1],sd.rw[1],0,1)
nom = posteriorp(var_star,tmu[2],p0,total[1],total[2],n0,x[1],x[2],0,q,f)
denom = posteriorp(tmu[1],tmu[2],p0,total[1],total[2],n0,x[1],x[2],0,q,f)
if (is.nan(nom)){ # error catch
nom=0;
}
if (nom == -Inf){ # error catch
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){ # error catch
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[1] <- accept[1]+1
tmu[1] <- var_star # accept proposed value
}
gibb.sample[i,1]=var_star
# # # # # # # # # # # # # # #
## p2 - p8 : marginal posteriors involve previous, current, and next p's
for (j in 2 : 8){
var_star <- rtnorm(1,tmu[j],sd.rw[j],0,1)
nom =
   posteriorp(var_star,tmu[j+1],tmu[j-1],total[j],total[j+1],total[j-1],x[j],x[j+1],0,q,
denom =
   posteriorp(tmu[j],tmu[j+1],tmu[j-1],total[j],total[j+1],total[j-1],x[j],x[j+1],0,q,f)
if (is.nan(nom)){
nom=0;
}
if (nom == -Inf){
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[j] <- accept[j]+1
tmu[j] <- var_star
}
gibb.sample[i,j]=var_star
}
# # # # # # # # # # # # # # #
```

```
## p9 : marginal posterior involves only p8 and p9
var_star <- rtnorm(1,tmu[9],sd.rw[9],0,1)
nom = posterior_individ(var_star,tmu[8],x[9],x[8],total[9],total[8],q,f)
denom = posterior_individ(tmu[9],tmu[8],x[9],x[8],total[9],total[8],q,f)
if (is.nan(nom)){
nom=0;
}
if (nom == -Inf){
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[9] <- accept[9]+1
tmu[9] <- var_star
}
gibb.sample[i,9]=var_star
# # # # # # # # # # # # # # #
# q : marginal posterior involves all priors
var_star= rtnorm(1,q,sd.rw[10],0)
nom=dgamma(var_star,10.254,10)# prior centered @ one
denom=dgamma(q,10.254,10)
nom = nom*posterior_q(tmu[1],p0,x[1],0,total[1],n0,var_star,f)
denom = denom*posterior_q(tmu[1],p0,x[1],0,total[1],n0,q,f)
for (k in 2:9){
nom=nom*posterior_q(tmu[k],tmu[k-1],x[k],x[k-1],total[k],total[k-1],var_star,f)
denom=denom*posterior_q(tmu[k],tmu[k-1],x[k],x[k-1],total[k],total[k-1],q,f)
}
if (is.nan(nom)){
nom=0;
}
if (nom == -Inf){
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[10] <- accept[10]+1
q <- var_star
```

```
}
gibb.sample[i,10]=var_star
}
accept/sample

burnin=2000
plot(gibb.sample[burnin:sample,1])
plot(gibb.sample[burnin:sample,2])
plot(gibb.sample[burnin:sample,3])
plot(gibb.sample[burnin:sample,4])
plot(gibb.sample[burnin:sample,5])
plot(gibb.sample[burnin:sample,6])
plot(gibb.sample[burnin:sample,7])
plot(gibb.sample[burnin:sample,8])
plot(gibb.sample[burnin:sample,9])
plot(gibb.sample[burnin:sample,10])

m1=mean(gibb.sample[burnin:sample,1])
m2=mean(gibb.sample[burnin:sample,2])
m3=mean(gibb.sample[burnin:sample,3])
m4=mean(gibb.sample[burnin:sample,4])
m5=mean(gibb.sample[burnin:sample,5])
m6=mean(gibb.sample[burnin:sample,6])
m7=mean(gibb.sample[burnin:sample,7])
m8=mean(gibb.sample[burnin:sample,8])
m9=mean(gibb.sample[burnin:sample,9])
m10=mean(gibb.sample[burnin:sample,10])

v1=var(gibb.sample[burnin:sample,1])
v2=var(gibb.sample[burnin:sample,2])
v3=var(gibb.sample[burnin:sample,3])
v4=var(gibb.sample[burnin:sample,4])
v5=var(gibb.sample[burnin:sample,5])
v6=var(gibb.sample[burnin:sample,6])
v7=var(gibb.sample[burnin:sample,7])
v8=var(gibb.sample[burnin:sample,8])
v9=var(gibb.sample[burnin:sample,9])
v10=var(gibb.sample[burnin:sample,10])

hist(gibb.sample[burnin:sample,10],ylim=c(0,1.75),prob=TRUE,xlab='',col=00a1c0,main=expr
   Histogram of q'[u]),cex=1)
hist(gibb.sample[burnin:sample,1],prob=TRUE,xlab='',main='Posterior
   Histogram of P(Undiagnosed), 2005')
hist(gibb.sample[burnin:sample,2],prob=TRUE,xlab='',main='Posterior
   Histogram of P(Undiagnosed), 2006')
```

```
hist(gibb.sample[burnin:sample,3],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2007')
hist(gibb.sample[burnin:sample,4],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2008')
hist(gibb.sample[burnin:sample,5],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2009')
hist(gibb.sample[burnin:sample,6],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2010')
hist(gibb.sample[burnin:sample,7],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2011')
hist(gibb.sample[burnin:sample,8],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2012')
hist(gibb.sample[burnin:sample,9],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2013')
```

Appendix C. Undiagnosed Population R Code

```
1
2  # # # # # # #
3  library(MCMCpack)
4  library(pracma)
5  library(msm)
6  library(rmutil)
7  # # # # # #  DATA  # # # # # #
8  total = c(896.9,923.2,951.5,979.7,1006.5,1032.6,1057.8,1082.1,1104.6)
9  undiagnosed = c(.218,.21,.201,.194,.187,.181,.176,.17,.164)
10 x=undiagnosed*total
11 plot(total)
12 plot(year,undiagnosed,xlab=Year,ylab=,xaxt='n',yaxt='n')
13 axis(1,at=year,labels=year)
14 axis(2,at=undiagnosed,labels=undiagnosed)
15 title(main=Estimated Undiagnosed HIV Population)
16 plot(x)
17 v=matrix(0,8)
18 for (i in 1:8){v[i]=undiagnosed[i+1]/undiagnosed[i]}
19 mean(v)
20 var(v)
21 # # # # # #
22 # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
23 # # # # # #
24
25 p0=.2 #expert prior - eqsual to proportion
26 n0=.25*total[1]
27
28 qs0=1 # assume no change:
29 beta0_qs=.5; alpha0_qs=2; #uninformative
30
31 qsprior=function(alpha,beta,q){
32 q^(alpha-1)*exp(-q/beta)
33 }
34
35 posteriorp=function(pt,pt1,pt.1,nt,nt1,nt.1,xt,xt1,xt.1,q,f){
36 ## this year t
37 talpha=pt.1*nt.1*f*q
38 tbeta=(1-pt.1*q)*nt.1*f
39
40 ## next year t+1
41 t1alpha=pt*nt*q*f
42 t1beta=(1-pt*q)*nt*f
```

```r
a=dbeta(pt,talpha,tbeta)*dbeta(pt,xt+1,nt-xt+1)
b=dbeta(pt1,t1alpha,t1beta)*dbeta(pt1,xt1+1,nt1-xt1+1)
a*b
}

posterior_individ=function(pt,pt.1,xt,xt.1,nt,nt.1,q,f){
talpha=pt.1*nt.1*f*q;  tbeta=(1-pt.1*q)*nt.1*f
dbeta(pt,xt+1,nt-xt+1)*dbeta(pt,talpha,tbeta)

}

posterior_q=function(pt,pt.1,xt,xt.1,nt,nt.1,q,f){
talpha=pt.1*nt.1*f*q;  tbeta=(1-pt.1*q)*nt.1*f
dbeta(pt,talpha,tbeta)
}

# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #
# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #
# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #

##### GIBBS SAMPLER #####
sample=100000
gibb.sample=matrix(0,sample,11) #columns = parametrs; check for convergence
tmu=matrix(.2,9,1)
q=.9
var_star = 0 # dummy variable for each iter

# # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #  # # # # # # # # # # # # # # #

sd.rw=matrix(.05,11);accept=matrix(0,11)
sd.rw[10]=.3

f=.1
for (i in 1:sample ){
set.seed(i)
# # # # # # # # # # # # # # #
## 1
var_star <- rtnorm(1,tmu[1],sd.rw[1],0,1)
nom = posteriorp(var_star,tmu[2],p0,total[1],total[2],n0,x[1],x[2],0,q,f)
denom = posteriorp(tmu[1],tmu[2],p0,total[1],total[2],n0,x[1],x[2],0,q,f)
if (is.nan(nom)){
```

```
nom=0;
}
if (nom == -Inf){
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[1] <- accept[1]+1
tmu[1] <- var_star
}
gibb.sample[i,1]=var_star
# # # # # # # # # # # # # # #
## 2 - 8
for (j in 2 : 8){
var_star <- rtnorm(1,tmu[j],sd.rw[j],0,1)
nom =
   posteriorp(var_star,tmu[j+1],tmu[j-1],total[j],total[j+1],total[j-1],x[j],x[j+1],0,q,
denom =
   posteriorp(tmu[j],tmu[j+1],tmu[j-1],total[j],total[j+1],total[j-1],x[j],x[j+1],0,q,f)
if (is.nan(nom)){
nom=0;
}
if (nom == -Inf){
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[j] <- accept[j]+1
tmu[j] <- var_star
}
gibb.sample[i,j]=var_star

}
# # # # # # # # # # # # # # #
## 9
var_star <- rtnorm(1,tmu[9],sd.rw[9],0,1)
nom = posterior_individ(var_star,tmu[8],x[9],x[8],total[9],total[8],q,f)
```

```
denom = posterior_individ(tmu[9],tmu[8],x[9],x[8],total[9],total[8],q,f)
if (is.nan(nom)){
nom=0;
}
if (nom == -Inf){
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[9] <- accept[9]+1
tmu[9] <- var_star
# likelihood[i,9]=nom;
}
gibb.sample[i,9]=var_star
# # # # # # # # # # # # # # #
# q
var_star= rtnorm(1,q,sd.rw[10],0)
#upper= q+sd.rw[10]; lower = q-sd.rw[10]
#if (lower<0){lower=0}
#var_star= runif(1,lower,upper)
nom=dgamma(var_star,9.788,10)
denom=dgamma(q,9.788,10)
nom = nom*posterior_q(tmu[1],p0,x[1],0,total[1],n0,var_star,f)
denom = denom*posterior_q(tmu[1],p0,x[1],0,total[1],n0,q,f)
for (k in 2:9){
nom=nom*posterior_q(tmu[k],tmu[k-1],x[k],x[k-1],total[k],total[k-1],var_star,f)
denom=denom*posterior_q(tmu[k],tmu[k-1],x[k],x[k-1],total[k],total[k-1],q,f)
}
if (is.nan(nom)){
nom=0;
}
if (nom == -Inf){
nom = 0;
}
alpha <- min(1, nom/denom)
if (is.nan(alpha)){
alpha=0
}
r <- runif(1,0,1)
if (r<=alpha ) {
accept[10] <- accept[10]+1
```

```
q <- var_star
}
gibb.sample[i,10]=var_star

}
accept/sample
mean(gibb.sample[burnin:sample,10])

burnin=2000
plot(gibb.sample[burnin:sample,1])
plot(gibb.sample[burnin:sample,2])
plot(gibb.sample[burnin:sample,3])
plot(gibb.sample[burnin:sample,4])
plot(gibb.sample[burnin:sample,5])
plot(gibb.sample[burnin:sample,6])
plot(gibb.sample[burnin:sample,7])
plot(gibb.sample[burnin:sample,8])
plot(gibb.sample[burnin:sample,9])
plot(gibb.sample[burnin:sample,10])
00FF00 55FF00 AAFF00 FFFF00
hist(gibb.sample[burnin:sample,10],ylim=c(0,1.75),prob=TRUE,xlab='',col=00a1c0,main=expr
    Histogram of q'[u]),cex=1)
curve(dgamma(x,9.79,10),xlim=c(0,10),lwd=6,col=b57786,add=TRUE,cex=.75)

hist(gibb.sample[burnin:sample,1],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2005')
hist(gibb.sample[burnin:sample,2],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2006')
hist(gibb.sample[burnin:sample,3],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2007')
hist(gibb.sample[burnin:sample,4],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2008')
hist(gibb.sample[burnin:sample,5],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2009')
hist(gibb.sample[burnin:sample,6],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2010')
hist(gibb.sample[burnin:sample,7],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2011')
hist(gibb.sample[burnin:sample,8],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2012')
hist(gibb.sample[burnin:sample,9],prob=TRUE,xlab='',main='Posterior
    Histogram of P(Undiagnosed), 2013')

m1=mean(gibb.sample[burnin:sample,1])
m2=mean(gibb.sample[burnin:sample,2])
```

```
m3=mean(gibb.sample[burnin:sample,3])
m4=mean(gibb.sample[burnin:sample,4])
m5=mean(gibb.sample[burnin:sample,5])
m6=mean(gibb.sample[burnin:sample,6])
m7=mean(gibb.sample[burnin:sample,7])
m8=mean(gibb.sample[burnin:sample,8])
m9=mean(gibb.sample[burnin:sample,9])
m10=mean(gibb.sample[burnin:sample,10])
means2=c(m1,m2,m3,m4,m5,m6,m7,m8,m9)
y=c(1,2,3,4,5,6,7,8,9);year=c(2005,2006,2007,2008,2009,2010,2011,2012,2013)
plot(year,means2,xlab='',ylab='',ylim=c(.1,.25),type=p,pch=21,cex=2,col=red)
boxplot(gibb.sample[burnin:sample,1:9],xlab='',ylab='',xaxt='n')
lines(year,undiagnosed,type=p,pch=22,cex=2,col=blue)
axis(1,at=1:9,labels=year)
arrows(year,
   means2-stdevs,year,means2+stdevs,length=0.05,angle=90,code=3,col=red)
axis(2,at=means2,labels=means2)
legend(2005,.155,c('Estimate','Observed'),pch =
   c(21,22),col=c('red','blue'))
title(main=Posterior Estimates of Undiagnosed Proportion)

plot(year,means2,xlab='',ylab='',ylim=c(.1,.25),type=p,pch=21,cex=2,col=red)
lines(year,undiagnosed,type=p,pch=22,cex=2,col=blue)
arrows(year,
   means2-stdevs,year,means2+stdevs,length=0.05,angle=90,code=3,col=red)
legend(2005,.155,c('Estimate','Observed'),pch =
   c(21,22),col=c('red','blue'))
title(main=Posterior Estimates of Undiagnosed Proportion)


v1=var(gibb.sample[burnin:sample,1])
v2=var(gibb.sample[burnin:sample,2])
v3=var(gibb.sample[burnin:sample,3])
v4=var(gibb.sample[burnin:sample,4])
v5=var(gibb.sample[burnin:sample,5])
v6=var(gibb.sample[burnin:sample,6])
v7=var(gibb.sample[burnin:sample,7])
v8=var(gibb.sample[burnin:sample,8])
v9=var(gibb.sample[burnin:sample,9])
v10=var(gibb.sample[burnin:sample,10])
plot(vars/n)
plot(c(v1,v2,v3,v4,v5,v6,v7,v8,v9))
vars=c(v1,v2,v3,v4,v5,v6,v7,v8,v9)
stdevs=sqrt(vars)
arrows(year, means2-stdevs,year,means2+stdevs,length=0.05,angle=90,code=3)
```

```
248
249 quantile((gibb.sample[burnin:sample,1]),c(.05,.95))
250 quantile((gibb.sample[burnin:sample,2]),c(.05,.95))
251 quantile((gibb.sample[burnin:sample,3]),c(.05,.95))
252 quantile((gibb.sample[burnin:sample,4]),c(.05,.95))
253 quantile((gibb.sample[burnin:sample,5]),c(.05,.95))
254 quantile((gibb.sample[burnin:sample,6]),c(.05,.95))
255 quantile((gibb.sample[burnin:sample,7]),c(.05,.95))
256 quantile((gibb.sample[burnin:sample,8]),c(.05,.95))
257 quantile((gibb.sample[burnin:sample,9]),c(.05,.95))
258 quantile((gibb.sample[burnin:sample,10]),c(.05,.95))
```

References

[1] Reinventing scientific talent | The NSF 2026 Idea Machine. Available at: https://nsf2026imgallery.skild.com/entries/reinventing-scientific-talent. (Accessed: 4th June 2019)

[2] Barone, L., Williams, J. & Micklos, D. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. PLoS Comput Biol 13, e1005755 (2017).

[3] Woodin, T., Carter, V.C., and Fletcher, L. (2010). Vision and change in biology undergraduate education, A Call for Action–Initial Responses. Cell Biology Education 9, 71–73.

[4] Chen, M.M., Scott, S.M., and Stevens, J.D. (2018). Technology as a tool in teaching quantitative biology at the secondary and undergraduate levels: a review. Letters in Biomathematics 5, 30–48.

[5] Vyshemirsky, V. & Girolami, M. BioBayes: A software package for Bayesian inference in systems biology. Bioinformatics 24, 1933–1934 (2008).

[6] Bois, F. Y. GNU MCSim: Bayesian statistical inference for SBML-coded systems biology models. Bioinformatics 25, 1453–1454 (2009).

[7] Warne, D. J., Baker, R. E. & Simpson, M. J. Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. J. R. Soc. Interface 16, 20180943 (2019).

[8] Golightly, A. & Wilkinson, D. J. Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics 61, 781–788 (2005).

[9] Hug, S. et al. High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. Mathematical Biosciences 246, 293–304 (2013).

[10] Argüello-Miranda, O., Liu, Y., Wood, N. E., Kositangool, P. & Doncic, A. Integration of multiple metabolic signals determines cell fate prior to commitment. Molecular Cell 71, 733-744.e11 (2018).

[11] Yeh, J. E., Toniolo, P. A., & Frank, D. A. (2013). JAK2-STAT5 signaling: A novel mechanism of resistance to targeted PI3K/mTOR inhibition. JAK-STAT, 2(4), e24635. doi:10.4161/jkst.24635

[12] Wilkinson, D. J. Bayesian methods in bioinformatics and computational systems biology. Briefings in Bioinformatics 8, 109–116 (2006).

[13] Bartlett, J. A. (2002). Addressing the challenges of adherence. JAIDS, 29(S1), S2-S10.

[14] Bangsberg, D. R., Acosta, E. P., Gupta, R., Guzman, D., Riley, E. D., Harrogan, P. R., Parkin, N., & Deeks, S. G. (2006). Adherence resistance relationships for protease and non-nucleoside reverse transcriptase inhibitors explained by virological fitness. AIDS, 20(2), 223–231.

[15] Paterson, D. L., Swindells, S., Mohr, J., Brester, M., Vergis, E. N., Aquier, C., Wagener, M. M., & Singh, N. (2000). Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. Annals of Internal Medicine, 133, 21–30.

[16] America's hidden H.I.V. epidemic [Internet]. [cited 2017 Jun 23]. Available from: https://www.nytimes.com/2017/06/06/magazine/americas-hidden-hiv-epidemic.html?nytmobile=0

[17] Chen, Y., Dale, R., He, H., Le, Q.-A.T., 2017. Posterior estimates of dynamic constants in HIV transmission modeling. Computational and Mathematical Methods in Medicine 2017, 1–8. https://doi.org/10.1155/2017/1093045

[18] Simoni JM, Huh D, Wilson IB, Shen J, Goggin K, Reynolds NR, et al. Racial/Ethnic disparities in ART adherence in the United States: Findings from the MACH14 study. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2012 Aug;60(5):466–72.

[19] Davey DJ, Beymer M, Roberts CP, Bolan RK, Klausner JD. Differences in risk behavior and demographic factors between men who have sex with men with acute and nonacute human immunodeficiency virus infection in a community-based testing program in Los Angeles. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2017;74(4):e97–e103.

[20] Velasco-Hernandez JX, Gershengorn HB, Blower SM. Could widespread use of combination antiretroviral therapy eradicate HIV epidemics? The Lancet infectious diseases. 2002;2(8):487–493.

[21] Lanoy E, Mary-Krause M, Tattevin P, Perbost I, Poizot-Martin I, Dupont C, et al. Frequency, determinants and consequences of delayed access to care for HIV infection in France. Antiviral therapy. 2007;12(1):89.

[22] Stein JA, Nyamathi A. Gender differences in behavioral and psychosocial predictors of HIV testing and return for test results in a high-risk population. AIDS Care. 2000;12(3):345–56.

[23] Pinkerton SD. HIV Transmission Rate Modeling: A Primer, Review, and Extension. AIDS and Behavior. 2012 May;16(4):791–6.

[24] Arroyo, M.J.H., Cabrera S.E., Correa, R.S., Merino, M.P.V., Gómez, A.I., Hurlé A.D. Impact of a pharmaceutical care program on clinical evolution and antiretroviral treatment adherence: a 5-year study. Patient Preference and Adherence. 2013 Aug;729.

[25] Bangsberg DR, Deeks SG. Is average adherence to HIV antiretroviral therapy enough? Journal of general internal medicine. 2002;17(10):812–813.

[26] Valdiserri RO, Forsyth AD, Yakovchenko V, Koh HK. Measuring what matters: development of standard HIV core indicators across the US Department of Health and Human Services [Internet]. SAGE Publications

Sage CA: Los Angeles, CA; 2013 [cited 2017 Aug 8]. Available from: http://journals.sagepub.com/doi/pdf/10.1177/003335491312800504

[27] Marks G, Crepaz N, Senterfitt JW, Janssen RS. Meta-analysis of high-risk sexual behavior in persons aware and unaware they are infected with HIV in the United States: implications for HIV prevention programs. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2005;39(4):446–453.

[28] Schneider MF, Gange SJ, Williams CM, Anastos K, Greenblatt RM, Kingsley L, et al. Patterns of the hazard of death after AIDS through the evolution of antiretroviral therapy: 1984–2004. Aids. 2005;19(17):2009–2018.

[29] Rasmussen DA, Volz EM, Koelle K. Phylodynamic inference for structured epidemiological models. PLoS computational biology. 2014;10(4):e1003570.

[30] Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, et al. Prevention of HIV-1 infection with early antiretroviral therapy. New England journal of medicine. 2011;365(6):493–505.

[31] Britton T. Stochastic epidemic models: A survey. Mathematical Biosciences. 2010 May;225(1):24–35.

[32] Sayles JN, Wong MD, Kinsler JJ, Martins D, Cunningham WE. The association of stigma with self-reported access to medical care and antiretroviral therapy adherence in persons living with HIV/AIDS. Journal of General Internal Medicine. 2009 Oct;24(10):1101–8.

[33] Montaner JS, Hogg R, Wood E, Kerr T, others. The case for expanding access to highly active antiretroviral therapy to curb the growth of the HIV epidemic. The Lancet. 2006;368(9534):531.

[34] Kinsler JJ, Wong MD, Sayles JN, Davis C, Cunningham WE. The effect of perceived stigma from a health care provider on access to care among a low-income HIV-positive population. AIDS Patient Care and STDs. 2007 Aug;21(8):584–92.

[35] Amico KR, Barta W, Konkle-Parker DJ, Fisher JD, Cornman DH, Shuper PA, et al. The information-motivation-behavioral skills model of ART adherence in a deep south HIV+ clinic sample. AIDS and Behavior. 2009 Feb;13(1):66–75.

[36] Granich RM, Gilks CF, Dye C, De Cock KM, Williams BG. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. The Lancet. 2009;373(9657):48–57.

[37] Song R, Hall HI, Green TA, Szwarcwald CL, Pantazis N. Using CD4 data to estimate HIV incidence, prevalence, and percent of undiagnosed infections in the United States. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2017;74(1):3–9.

[38] Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas, 2014. HIV Surveillance Supplemental Report 2016;21(No. 4). http://www.cdc.gov/hiv/library/reports/surveillance/. Published July 2016. Accessed 3/14/2017.

[39] Centers for Disease Control and Prevention. Behavioral and clinical characteristics of persons receiving medical care for HIV infection—medical monitoring project, United States, 2014 Cycle (June 2014–May 2015). HIV Surveillance Special Report 17.

[40] Vlahov D, Graham N, Hoover D, Flynn C, Bartlett JG, Margolick JB, et al. Prognostic indicators for AIDS and infectious disease death in HIV-infected injection drug users: plasma viral load and CD4+ cell count. JAMA. 1998 Jan 7;279(1):35–40.

[41] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[42] Bruce Swihart and Jim Lindsey (2017). Rmutil: Utilities for nonlinear regression and repeated measurements models. R package version 1.1.0. https://CRAN.R-project.org/package=rmutil

[43] MATLAB R2016b, The MathWorks, Inc., Natick, Massachusetts, United States.

[44] Michigan Department of Health and Human Services. Annual HIV Surveillance Report, City of Detroit. July 2015. http://www.michigan.gov/documents/mdch/Detroit_496828_7.pdf

Vita

Renee Dale obtained a bachelor's in biology and philosophy from LSU in 2013. After taking a class on mathematical modeling in her senior year, she entered grad school studying mathematical biology, eventually completing a master's in 2015, and her PhD in 2019. During this time, she was inspired to study statistics after managing to pass Dr Escobar's probability courses and learning about Bayesian statistics. Thanks to this, she was able to learn about data mining and regression, and how to do some cool stuff. Despite pandemic delays, she plans to receive her master's in Statistics in May 2021.