Mathematical Modeling, R_0, and the Importance of Uncertainty

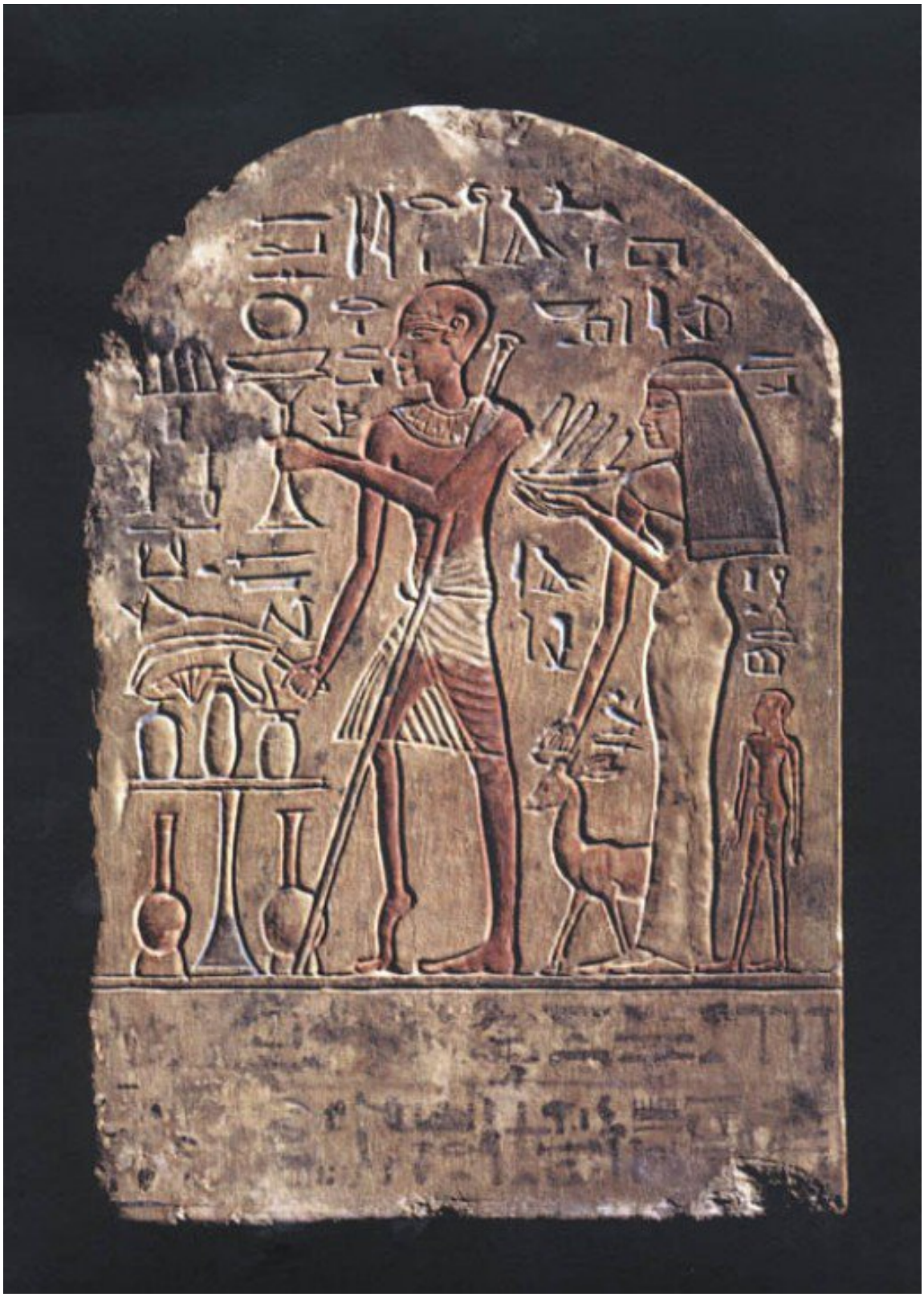
by

Christopher James Henry

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Epidemiological Science)
in The University of Michigan
2017

Doctoral Committee:

Assistant Professor Marisa C. Eisenberg, Co-Chair
Professor James S. Koopman, Co-Chair
Professor Edward L. Ionides
Associate Professor Rafael Meza
Professor Mark E.J. Newman



Christopher J. Henry

chrishen@umich.edu

ORCID iD: 0000-0001-5621-708X

ACKNOWLEDGEMENTS

Many thanks to my advisor, Dr. James S. Koopman, who has been a valuable source of advice and encouragement throughout my doctoral program, in addition to being the one who encouraged me to pursue a Ph.D. in Epidemiology in the first place.

Many thanks also to Dr. Michael Hayashi, who as a postdoc in the Epidemiology department provided moral support, encouragement, advice, directions to the LaTeX template used in this dissertation, and, when necessary, intense prodding. Without his particular combination of sternness and beneficence, this dissertation would have taken far longer to complete.

Thanks also to Dr. Jin Ji, who authored that LaTeX template and made it available for the use of others, to those who subsequently developed it further, and to the Biostatistics department, for hosting it.

Deep thanks also to my parents, who have always been supportive of my academic pursuits, both financially and emotionally, the latter both gently and not-so-gently, as the situation has demanded. Thanks to my friends, who have listened to me both babble excitedly about cool aspects of my research and moan at length about the process of writing it up. And thanks to everyone else whom I have neglected to mention, but without whom, this dissertation would not have possible.

PREFACE

Chapter II has been published with Xinyu Zhang, Lin Zhong, Ethan Romero-Severson, Shah Jamal Alam, Erik M. Volz, and James S. Koopman in Statistical Communications in Infectious Disease[132]. Chapter III and Appendix A have been published with James S. Koopman in Scientific Reports[58]. Chapters IV and V have been published with James S. Koopman, JoonHa Park, Marisa C. Eisenberg, Edward L. Ionides, and Joseph N. Eisenberg in Epidemics[77]. A version of Chapter VI will be submitted for publication, with the following authors: Christopher J. Henry, Michael A.L. Hayashi, Marisa C. Eisenberg, and James S. Koopman.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Mathematical modeling of infectious disease transmission dynamics can never represent all features of those dynamics. Its goal must be to represent all those features which are salient for the purpose of making particular inferences that are relevant to public health decisions. The way to identify these features is through an inference robustness assessment, a continuous cycle in which we realistically relax simplifying assumptions, determine whether our inferences are robust to that relaxation, and then, based on the answer to that question, either gather additional data or relax additional assumptions.

In this dissertation, I examine the consequences of realistically relaxing particular simplifying assumptions about the transmission dynamics of two very different pathogens, HIV and poliovirus. I show that key inferences about these systems are not robust to such relaxation, and highlight some ways in which additional data might be gathered in order to constrain the plausible parameter space. Chapters II and III of this dissertation deal with the effects of episodic risk in HIV transmission systems; Chapters IV through VI cover the effects of waning in poliovirus transmission systems.

The per-act probability of sexual transmission of HIV is especially high during early HIV infection (EHI). Any behavioral pattern that increases the expected number of sex acts early in HIV infection will naturally interact with this. One such pattern is episodic risk. In this dissertation, I examine how episodic risk can greatly affect both the role of EHI in sustaining HIV transmission and the relative ease of controlling transmission through Universal Test and Treat measures. Inferences

made on the basis of observations about the degree of risk heterogeneity at a given point in time are shown not to be robust to uncertainty in the degree to which risk heterogeneity is episodic, rather than static.

Only a minute fraction of poliovirus infections result in the acute flacid paralysis (AFP) for which poliovirus is notorious. Nevertheless, AFP surveillance remains the backbone of efforts to detect circulating poliovirus. Consequently, the circulation of poliovirus in the absence of paralytic cases may be said to be "silent" with respect to our primary means of detecting it. Silent circulation that is sustained beyond the Global Polio Eradication Initiative's standard of 3 years without a detected polio case in order to certify a country polio-free has the potential to seriously jeopardize eradication efforts.

The potential for such sustained silent circulation is increased by a high prevalence of partial immunity to poliovirus infection. Even very low levels of immunity are sufficient to effectively eliminate the risk of developing paralysis upon exposure to virulent strains of poliovirus. There is substantial evidence that even full immunity can wane sufficiently over time to make (non-paralytic) infection possible again. Although the importance of including waning in models of poliovirus transmission is increasingly recognized by prominent modelers in the field, this inclusion has often failed to take adequate account of our level of uncertainty as to the speed and depth of that waning. In this dissertation, I present evidence that important inferences about the potential for sustained silent circulation are not robust to that uncertainty.

On a methodological note, in Appendix A, I present an effective reproduction number analogue to the Next-Generation Matrix basic reproduction number. In Chapter VI, I present a hybrid deterministic-stochastic approach for efficient simulation of dynamics during silent circulation.

CHAPTER I

Introduction

Mathematical modeling of transmission dynamics offers great opportunities for understanding current disease dynamics, making predictions of future dynamics, and making policy recommendations. But no mathematical model can ever be a complete representation of all aspects of transmission dynamics of a particular disease, for three distinct reasons. First, we may be wholly ignorant of some relevant aspect or aspects of human behavior, pathogen characteristics, or both. As we improve our methodology, this becomes less common, but it can never be eliminated entirely, especially as both society and pathogens are continually changing. Second, we may recognize the existence of a phenomenon, but lack certainty regarding the numbers that characterize it quantitatively. For example, we might be aware that enteric infection with non-polio enteroviruses at the time of oral polio vaccine (OPV) administration reduces the probability of successful vaccination[93], but be uncertain as to how severe that reduction is.

Finally, we may deliberately omit aspects of reality in order to simplify our models. In some cases, this may be out of necessity, due to limitations in available computational resources that constrain how complex of a model we can have and still be able to make predictions in a timely enough fashion for them to be of use. In other cases,

we may deliberately simplify a model in order to obtain a better understanding of the processes that give rise to certain patterns, or to attempt to make the model more robust to errors in other assumptions we may be forced to make.

These limitations, while sobering, are not insurmountable. In the famous words of George Box, "All models are wrong, but some are useful."[16] Our key task, therefore, is to analyze when omitting details or being uncertain about the parameters that characterize them actually threatens the concrete inferences that we can make about important real-world outcomes, and when it does not. In order to do this, we can examine the simplifying assumptions that are present in a model, realistically relax them, and see whether the inferences we care about change as a result. This process, when iterated and combined with the gathering of data in order to narrow the range of plausible parameter space when our inferences of interest differ across that space, constitutes an *inference robustness assessment.*[75]

This dissertation aims to realistically relax simplifying assumptions with respect to certain aspects of the transmission dynamics of two very different pathogens, HIV and poliovirus. Despite their many differences, these two have one salient feature in common: Most infections are not detected until long after they have occurred, if at all. Because of this, direct observation of who is transmitting to whom is impossible in the vast majority of cases. The consequences of uncertainty about the importance of various factors in transmission dynamics that this causes are central to this dissertation.

1.1 Background regarding HIV

Over the past several decades, there has been great progress in controlling HIV. Improvements in antiretroviral treatment have saved millions of lives. But HIV

incidence among men who have sex with men (MSM) continues to rise, indicating the need for continuing improvement in transmission control strategies.[102]

One of several major strategies that has been put forward and implemented in order to control the transmission of HIV is Treatment as Prevention, or TasP. Antiretroviral treatment not only drastically slows the progression of HIV infection, extending life expectancy and improving quality of life, it also, by reducing viral load in blood and other bodily fluids, reduces the probability of transmitting the virus to others. Thus, treating existing cases of HIV has the potential to prevent future cases. But in order for this to occur, existing cases must first be diagnosed[36].

Almost all HIV infections in the United States will eventually be diagnosed. However, most of these infections are not being detected until long after they begin[57]. This produces both an inability to observe HIV transmission patterns directly, and the elapsing of substantial time during which the infected individual cannot be treated, and therefore retains their full ability to make secondary transmissions. The importance of this fact is increased by the fact that the first 6 months or so[111] of HIV infection are characterized by a heavier viral load and a higher probability of transmission per sex act than the multi-year (even if untreated) chronic stage that follows (prior to a progression to pre-AIDS and AIDS, when the viral load and probability of transmission rise again). Because we cannot observe transmission patterns directly, we cannot count directly how many transmissions are occurring during this highly infectious but brief early stage, known as early HIV infection (EHI). However, evidence derived from the use of viral RNA sequence data suggests that this fraction is likely quite high[17]. There are multiple possible explanations for this, both biological (including, but not limited to, the aforementioned increased viral load) and social.

One possible social explanation which has received relatively little attention is changes in the riskiness of an individual's sexual behavior over time, which can be conceptualized in categorical terms as episodic risk, or in continuous terms as risk volatility[110]. It is widely recognized that individuals within the same broad transmission category can vary widely in the frequency and riskiness of the sex acts they engage in. This between-individual risk heterogenity has long been recognized as being important in HIV transmission dynamics, and in developing strategies for HIV control[85]. But attention to the role of within-individual variation in propensity to engage in risky behavior over time (episodic risk) has lagged far behind.

In the first two papers incorporated into this dissertation (the first as Chapter II, the second as Chapter III and Appendix A), I examine how episodic risk can greatly affect both (1) the role of acute HIV infection (AHI, a subphase of EHI[111]) and/or EHI more generally in sustaining HIV transmission and (2) the relative ease of controlling transmission through Universal Test and Treat (UT&T) measures. Inferences made on the basis of observations about the degree of risk heterogeneity at a given point in time are shown not to be robust to uncertainty about the degree to which risk heterogeneity is episodic, rather than static.

1.2 Background regarding Polio

In contrast to HIV, the vast majority of poliovirus infections will never be detected at all. This is because poliovirus is fundamentally a gastrointestinal pathogen that is spread through the fecal-oral route and whose primary symptom, in the vast majority of cases, is mild diarrhea. Even in cases where a more systemic infection occurs, the usual outcome is merely non-specific ("flu-like") symptoms. Only a minute fraction of infections — depending on poliovirus strain, between 1 in 200 and 1 in 2000, in

fully susceptible individuals — result in the acute flacid paralysis for which poliovirus is notorious[115].

Nevertheless, acute flacid paralysis (AFP) surveillance remains the backbone of efforts to detect the presence of circulating poliovirus worldwide[115]. Environmental surveillance (ES), which looks for the presence of poliovirus in sewage or intensely fecally-contaminated surface water, is gradually assuming increasing importance as a secondary method[46]. But it is still limited in its distribution, and is more difficult to implement in areas that lack sewage systems — precisely those areas where poor sanitation is most likely to result in a high potential for ongoing poliovirus transmission. Although some progress has been made in this area, far more is required. Consequently, the circulation of poliovirus in the absence of paralytic cases may be said to be "silent" with respect to our primary means of detecting it.

The potential for sustained silent circulation is particularly important because we are reaching the endgame for global polio eradication efforts, and the Global Polio Eradication Initiative (GPEI) has a standard of 3 years without a detected polio case in order to certify a country polio-free[87]. Making such a certification in error has the potential to result in decisions that could seriously jeopardize eradication efforts. The potential for such error can never be eliminated entirely. But even keeping it to an acceptably low level requires understanding the dynamics that make it more or less likely.

The potential for sustained silent circulation may be greatly enhanced by a high prevalence of partial immunity to poliovirus infection. Even very low levels of immunity are sufficient to drastically reduce the probability of developing paralysis upon exposure to virulent strains of poliovirus, even compared to the low baseline probability for fully susceptible individuals, to the point where it is a reasonable

approximation for most practical purposes to say that partial immunity is fully protective against paralysis. The most well-known source of such partial immunity is vaccination with injectable polio vaccine (IPV), a killed-virus vaccine. Another is partial immunity following an inadequate number of doses of oral polio vaccine (OPV), a live attenuated vaccine that requires 3 or more doses to provide full protection in high-transmission settings. But there is also substantial evidence that even full immunity, following adequate vaccination with OPV, can wane sufficiently over time to make (non-paralytic) infection possible again.

Although the importance of including waning in models of poliovirus transmission is increasingly recognized by prominent modelers in the field, this inclusion has often failed to take adequate account of our level of uncertainty as to the speed and depth of that waning. Consequently, if important inferences about the potential for sustained silent circulation are not robust to that uncertainty, then it is crucial to gather additional data to address that uncertainty. In this dissertation, I present evidence that that is in fact the case.

The urgency of completing global polio eradication, and hence of addressing any uncertainties regarding polio transmission dynamics that threaten that effort, is underscored by the recurring detection of "orphan" viruses whose genetic distance from other detected viruses indicates extended periods circulating undetected[55, 83]. In particular, such viruses were responsible for the reccurence, in 2016, of WPV1-caused acute flaccid paralysis (AFP) cases in Nigeria two years after WPV1 had last been detected in that country[83].

1.3 Brief discussion of transmission modeling

1.3.1 Stochastic and deterministic approaches

Disease transmission processes involve a great deal of chance. An individual who is infected with a particular pathogen might recover (or die) quickly, or slowly. If he is still alive, and still contagious, he might encounter another, uninfected, individual who lives in the same city on a given day, or not. If they encounter each other, the nature of that encounter might create a potential for transmission, or not. If that potential exists, the pathogen might successfully colonize the uninfected individual, or not. And so forth.

Some of these individual-level processes may well technically be deterministic, but only with information so detailed that we can never in practice obtain it. Others might well be truly random, unpredictable even in theory until they occur (at least under some interpretations of quantum mechanics). In any event, they are all *effectively* random from the perspective of an observing epidemiologist: Regardless of whether it is truly, mechanistically random whether or not a susceptible individual's encounter with an infectious individual's sneeze will result in an influenza infection, or mechanistically deterministic, but only with knowledge of the exact position of virions within the aerosolized water droplets, is relevant perhaps to philosophers, but not to transmission modelers.

There are three basic ways to incorporate this reality into our models of transmission dynamics. The first is to assign a probability to each possible outcome of a particular class of random event (such as whether a single instance of a particular sex act between an infected and a suscptible individual transmits HIV), and to calculate algebraically the exact probability of each outcome of interest(such as the number of new HIV infections occuring within the next five years following a possible

intervention).

Although a number of useful results have been obtained in this fashion, the resulting systems of equations are generally impossible to solve algebraically for all but the simplest systems. One solution to this problem is to simulate a (random) trajectory of the system numerous times, and to empirically approximate the probability of a given outcome by the fraction of simulations which generate that outcome. This approach, stochastic simulation, has the advantage of realism, but the disadvantage of computational complexity, which becomes more severe as the number of individuals in the population increases.

The other approach, a large-population approximation[27], is to replace discrete individuals with continuous (sub)population sizes, and to replace discrete *probabilities* of events occuring with continuous *rates* at which those events occur. Thus the trajectory of the system becomes the solution to a set of ordinary differential equations. This makes population-level outcomes deterministic, which greatly simplifies calculation. In many cases, the lack of random noise can also facilitate distinguishing qualitative changes in behavior with changing parameters (phase transitions) from merely quantitative changes. The downside is that we must be careful that these approximations are realistic. Such a *deterministic compartmental model* can allow transmission to continue for years — or even forever — after the point at which the number of individuals infected has fallen to a miniscule fraction of a single individual. In the real world, individuals are discrete, and when a deterministic approximation indicates an infected population of <1 individual for an extended period of time, this generally corresponds to *stochastic elimination*.

These two approaches, then, are fundementally complementary: Stochastic simulations work well for small populations, but become computationally difficult for

large ones. Deterministic compartmental models work well for large populations, but become unrealistic for small ones.

In this dissertation, I use both approaches as appropriate. When dealing with the HIV epidemic among American men who have sex wth men (MSM), relevant numbers are large, and so consistent use of deterministic compartmental models is reasonable. When dealing with polio eradication efforts, I begin by using a deterministic compartmental model (in Chapters IV and V), and then subsequently adopt a novel hybrid approach (in Chapter VI) that seeks to take advantage of the computational efficiency of deterministic compartmental models for dealing with large total populations, while also gaining the realism of stochastic models when dealing with small *infected* subpopulations during silent circulation.

1.3.2 The SIR model and its variants

The usual starting point of transmission modeling, both in a conceptual and in a historical sense, is the SIR compartmental model of Kermack and McKendrick[71]. In the simplest form of this model, individuals exist in one of three states: Susceptible (S), Infectious (I), and Recovered (R). In the basic model, transitions occur in a single direction: Individuals are born (or otherwise enter the population of interest; e.g., in a simulation of sexual transmission of HIV, "birth" may actually represent sexual debut) into the Susceptible state, transition to the Infectious state state upon infection, and transition to the Recovered state upon recovery. Death may occur from any of the three states.

1.3.2.1 Deterministic compartmental models

A deterministic compartmental version of the SIR model may be expressed by the following set of equations:

$$\frac{d}{dt}(S) = \mu N - \mu S - \frac{\chi \beta SI}{N}$$
$$\frac{d}{dt}(I) = -\frac{\chi \beta SI}{N} - \mu I + -\gamma I$$
$$\frac{d}{dt}(R) = \gamma I - \mu R$$

where μ is the mortality rate, χ is the contact rate, β is the per-contact transmission probability, γ is the recovery rate, and N is the total population size at equilibrium.

Numerous variations exist, based on the details of transmission dynamics for a given disease. For HIV, for example, infection is lifelong, and so the Recovered state does not exist. This makes an SI model more appropriate than an SIR model. For polio, infection is relatively brief, and is followed by strong immunity, but subsequent immune waning returns individuals to (partial) susceptibility. Consequently, a (modified) SIRS model is more appropriate than a simple SIR model. The models in this dissertation include these features and others, which will be dealt with in more detail in their respective chapters.

1.3.2.2 Stochastic models

For a stochastic model, we replace first-order deterministic flows of population between compartments with exponentially-distributed waiting times for individuals to transition between the corresponding states. Thus, each deterministic rate of flow between compartments is replaced by a rate parameter, having the same numerical value, that governs the distribution of waiting times before an individual in the state corresponding to the source compartment will transition to the state corresponding

to the sink compartment (if they do not leave the former state in some other fashion first). From this principle, and the memoryless property of the exponential distribution, we can obtain Gillespie's First-Reaction method[45]: At each step of the simulation, for each type of possible transition from one state to another, we first calculate rate parameters (known in this context as propensities). For the deterministic SIR model in the previous section, the corresponding stochastic propensities are:

$$\begin{aligned}
a_{(S,I,R)\to(S+1,I,R)} &= \mu N \\
a_{(S,I,R)\to(S-1,I,R)} &= \mu S \\
a_{(S,I,R)\to(S,I-1,R)} &= \mu I \\
a_{(S,I,R)\to(S,I,R-1)} &= \mu R \\
a_{(S,I,R)\to(S-1,I+1,R)} &= \frac{\chi\beta SI}{N} \\
a_{(S,I,R)\to(S,I-1,R+1)} &= \gamma I
\end{aligned}$$

Using these propensities, we then draw a random waiting time for each of these transitions:

$$t_i \sim \text{Exponential}(a_i)$$

And, finally, we select the transition with the lowest waiting time, increment the current time by that waiting time, and implement that transition. For example, if, for values drawn at time t_0, $t_{(S,I,R)\to(S,I-1,R)} < t_i \forall i \neq (S,I,R) \to (S,I-1,R)$, then at time $t_0 + t_{(S,I,R)\to(S,I-1,R)}$, the state of the system will shift to $(S,I-1,R)$.

From the mathematical properties of exponential distributions, it is possible to obtain an equivalent, but slightly more computationally efficient form of this algorithm, known as Gillespie's direct method[45]: We calculate the propensities in the

same way as before, but instead of calculating a waiting time for each propensity, we sum all of the propensities to obtain a waiting time until the next transition of any sort:

$$t_{\text{any}} \sim \text{Exponential}\,(\Sigma_i a_i)$$

and, separately, choose which transition occurs in proportion to their respective propensities:

$$P(\text{next transition is } k) = \frac{a_k}{\Sigma_i a_i}$$

This approach is used for the stochastic component of the hybrid deterministic-stochastic model in Chapter VI.

1.4 The basic reproduction number, R_0

One of the most fundamental concepts in the mathematical modeling of infectious disease transmission systems is the basic reproduction number (R_0). In an idealized, fully homogenous system, the basic reproduction number represents the average number of secondary transmissions that an infected individual will make over the course of their infection, assuming that everyone they make contact with is fully susceptible. Hence, in a fully susceptible population into which a sufficiently small number of infected individuals have been introduced, it represents the expected ratio of sizes of subsequent generations of infections. Therefore, if it is >1, an epidemic of non-trivial size will occur (with certainty in a deterministic compartmental model, and with non-zero probability in a stochastic model). If it is ≤ 1, and the susceptible population is sufficiently large, then introduced pathogen will (almost certainly, in the stochastic case) die out without generating such an epidemic. In the simple SIR model discussed in the previous section, the basic reproduction number is given by

the equation:

$$R_0 = \frac{\chi\beta}{\gamma + \mu}$$

When individuals do not all have the same contact rates, the mean contact rate for the population as a whole cannot be used in the above equation, because individuals with higher contact rates have both a higher propensity to become infected and a higher propensity, while infected, to infect others. When this is accounted for, the equation becomes[85]:

$$R_0 = \frac{\mathrm{E}\left(\chi\right)\beta}{\gamma + \mu}\left(1 + c_{\mathrm{V}}\left(\chi\right)^2\right)$$

In more complicated systems, the usual approach is to construct a Next-Generation Matrix, and to calculate its dominant eigenvalue[25]. This amounts to weighting individuals by their relative numbers in a given generation of infections under conditions of (proportional) exponential growth. This approach is discussed at much greater length in Chapter III and Appendix A. A modified version, pertaining to the effective reproduction number (R_{eff}) is developed and used in Chapter V.

1.5 Organization and aims of this dissertation

Chapters II and III deal with the effects of episodic risk in HIV transmission systems; Chapters IV through VI cover the effects of waning in poliovirus transmission systems.

In Chapters II and III, I present a simplified, deterministic model of episodic risk in men who have sex with men (MSM). Chapter II develops the model itself, and shows how multiple biological and behavioral factors have complex and non-monotonic effects on both the fraction of transmissions from acute infection and the endemic prevalence when the average transmission probability per sex act over the

course of infection is held fixed. Chapter III adds treatment to the model, in order to focus in on the potential for control of transmission through universal test and treat (UT&T). In this chapter, I extend the classic results presented in the equation above, regarding the effect of risk heterogeneity on the basic reproduction number (R_0) to consider the effect of the extent to which that risk heterogeneity is episodic rather than static. From this, I obtain the counterintuitive result that episodic risk can in many cases make control of transmission through universal test and treat easier, despite increasing the fraction of transmissions that come from early infection.

In Chapters IV and V, I present a deterministic model of poliovirus transmission that includes strain-specific characteristics, variable speed and depth of waning, and different histories of eradication efforts. I show that uncertainty about waning dynamics translates to uncertainty about the potential for silent circulation lasting longer than 3 years, and discuss ways that this uncertainty might be addressed. Chapter IV focuses on the general outline of the model and broad conclusions, while Chapter V goes into greater detail on how the model is constructed, how results were obtained, and the effects on our inferences of relaxing some of our own simplifying assumptions.

Finally, in Chapter VI, I present a hybrid deterministic-stochastic model that combines the strengths of the (deterministic) model from Chapters IV and V with those of a stochastic model[41] designed to estimate the probability that a given strain of poliovirus has been eliminated (i.e., that silent circulation has ended) at a given point in time, conditional on not having observed a return of detected paralytic cases. This allows me to replace the single outcomes of the purely deterministic model with probability distributions, in a more computationally efficient way than using a purely stochastic model would allow. Using this hybrid approach, I show that

for large areas of plausible parameter space, the median length of silent circulation is short enough to be consistent with the empirical experience that has guided the formulation of the GPEI's three-year rule. At the same time, for a large subset of these areas, the probability of much longer silent circulation (while less than 50%) is far too large to constitute an acceptable risk going forward.

CHAPTER II

Episodic HIV Risk Behavior Can Greatly Amplify HIV Prevalence and the Fraction of Transmissions from Acute HIV Infection

2.1 Chapter preface

The contents of this chapter are adapted from a paper[132] originally published in Statistical Communications in Infectious Disease.

In this chapter, I present and analyze a simplified, deterministic model of episodic risk in men who have sex with men (MSM). This model includes multiple risk groups, assortative mixing (conceptualized in terms of multiple mixing sites), and a simplified version of the natural history of HIV infection consisting of a short, highly-infectious acute phase (acute HIV infection, or AHI) and a longer, less-infectious chronic phase. The degree to which risk heterogeneity is episodic is parameterized in terms of the rate of transition from one risk group to the other. Further analysis, after the contents of this chapter were first published, revealed that there is a fully equivalent parameterization that is more mathematically and conceptually elegant. That parameterization is used in Chapter III. But in this chapter, I have kept the original parameterization for consistency with the previously published version.

By the same token, I have left intact the unfortunate choice to use χ and X to denote two unrelated quantities, namely the average contact rate and the fraction

of total (homogeneous) transmission potential that comes from the acute stage of infection, respectively.

Using this model, I examine how episodic risk can greatly affect both the role of AHI in sustaining HIV transmission and the relative ease of controlling transmission through Universal Test and Treat (UT&T) measures. I show that inferences made on the basis of observations about the degree of risk heterogeneity at a given point in time are not to be robust to uncertainty in the degree to which risk heterogeneity is episodic, rather than static.

2.2 Introduction

In the decades from its emergence, great strides have been made in controlling HIV. Antiretroviral treatment has greatly improved and has saved millions of lives. Neonatal transmission and transmission via needles among injecting drug users have been markedly reduced[56]. On the other hand, HIV incidence among men who have sex with men (MSM) continues to rise with increasing rates of treatment having little effect on controlling transmission in this group[56, 102]. There are diverse speculations as to why this might be the case including rapid evolution of the virus itself, safe-sex exhaustion, or emergence of versatility in terms of insertive or receptive sex roles among MSM[14, 15, 50, 126]. Nonetheless, convincing evidence and arguments as to how different factors contribute to our failure to control HIV transmission among MSM are lacking. The formulations we present here constitute a theory expressing one mechanism through which biological and social factors interact to maintain high infection levels in MSM populations.

Our ability to understand HIV transmission dynamics is hindered by the fact that we cannot observe HIV transmission patterns directly. In part, this is due

to most infections not being detected until long after transmission. That makes it impossible to trace all sexual partners and count directly how many transmissions are occurring during the highly infectious but brief acute HIV infection stage (AHI). It also makes it rare that AHI outbreaks are detected and investigated at the time they are occurring. Currently the existence of such outbreaks is detected through genetic clustering (c.f. [18, 19]) and phylogenetic relationships[63, 82]. As many as 64% of infections detected in the first 6–24 months after infection can be identified as part of genetic clusters[17, 18]. Quite possibly most clustering involves transmission during AHI but inferring AHI transmission fractions from genetic data require better theory. This chapter takes a step in that direction by developing a model that elucidates an important mechanism by which acute infections are linked into outbreaks.

The ability to use genetic data to test transmission theory has recently been advanced by new methods for incorporating transmission models into coalescent models[105, 129, 127]. Being able to fit different models to genetic data creates the potential to assess the robustness of inferences about the fraction of transmissions from AHI by realistically relaxing diverse aspects of transmission models[76]. These include partnership formation and duration patterns[72], specific behaviors such as anal or oral sex with or without protection, insertive and receptive behavior patterns[5], different sexual mixing sites[79], and heterogeneous sexual behavior in the population[7, 59]. In this chapter, we focus on elaboration of a particular model aspect that has especially strong effects: episodic risk behavior. Another paper published in the same symposium that the contents of this chapter were originally published in (as [132]) illustrated short-term temporal heterogeneity in sexual contact rates in a prospective risk behavior study designed to estimate transmission risks for different types of sexual contacts between MSM[108].

Episodic risk is defined by brief periods of higher risk behavior, possibly precipitated by random events such as a change in relationship status, drug use, or as an endogenous aspect of sexual behavior in general. We modeled episodic risk with flows back and forth between high and low contact rate states. The first paper to do this was published in 1997[78]. However, the predominant feeling has been that previous demonstration of high AHI transmission[65, 78] was only applicable during the early stages of the epidemic[104]. That is, once infection levels reach an endemic equilibrium, the fraction of transmissions is a simple function of the average duration of different stages and the transmission probabilities during those stages[2, 61].

In this chapter, we show how episodic risk can amplify AHI transmissions even after endemic equilibrium has been reached. Specifically, we show how behavioral changes in individuals over time can amplify transmission at the population level by amplifying transmission from acute HIV infection. The mechanisms include both increasing the contact rates of recently infected individuals and increasing the number of high risk susceptible individuals with whom high risk infected individuals can make contact.

2.3 Methods

2.3.1 Model formulation

We developed a deterministic compartmental model of HIV transmission in men who have sex with men that integrates episodic risk behavior, non-random mixing, and multiple stages of infection. The model has only two risk phases with different contact rates (high and low), two infection stages (acute and chronic), and two sexual mixing sites. There is a common mixing site where individuals in the high and low risk phases mix proportionally and a high-risk mixing site where only the high-risk individuals make contacts. Contacts are modeled as instantaneous and symmetrical

partnerships with an undefined number of sex acts having no specified direction or pattern (such as insertive or receptive, anal or oral).

The deterministic compartmental model is specified as a system of six differential equations for a MSM population, one for each compartment in the model. A schematic of the model is shown in Figure 2.1. The characters S, A and C for each compartment refer to the susceptible, acute HIV and chronic HIV stages respectively. The subscripts H and L refer to the high and low contact rate for each subpopulation respectively. The grey area in Figure 2.1 depicts the high-risk mixing site where individuals with high contact rate make a proportion of their contacts (ν) exclusively with each other. The dotted area designated as the common mixing site represents the site where individuals with high contact rate make the remaining fraction of their contacts ($1 - \nu$) with individuals with low contact rate proportionally. The parameters that govern the rate of each flow are shown in Figure 2.1 near the arrow depicting that flow. The values of parameters used in Figure 2.1 are given in Table 2.1.

Arrows between subpopulations with identical infection status represent the flow of individuals between these subpopulations due to the movement between high and low-risk phases. Φ_{H} represents the rate of turnover from high to low contact rate while Φ_{L} refers to the flow in the opposite direction. We assume a constant birth rate indicated by the arrows in Figure 2.1 pointing into S_{L} and S_{H}. Individuals leave the system either by a process of "natural removal" that accounts for cessation for sexual activity for any reason (horizontal arrows pointing out of each compartment in Figure 2.1) or by death of chronically infected individuals due to AIDS (vertical arrows pointing out of the C_{L} and C_{H} compartments). Table 2.1 presents a complete list of the model's parameters.

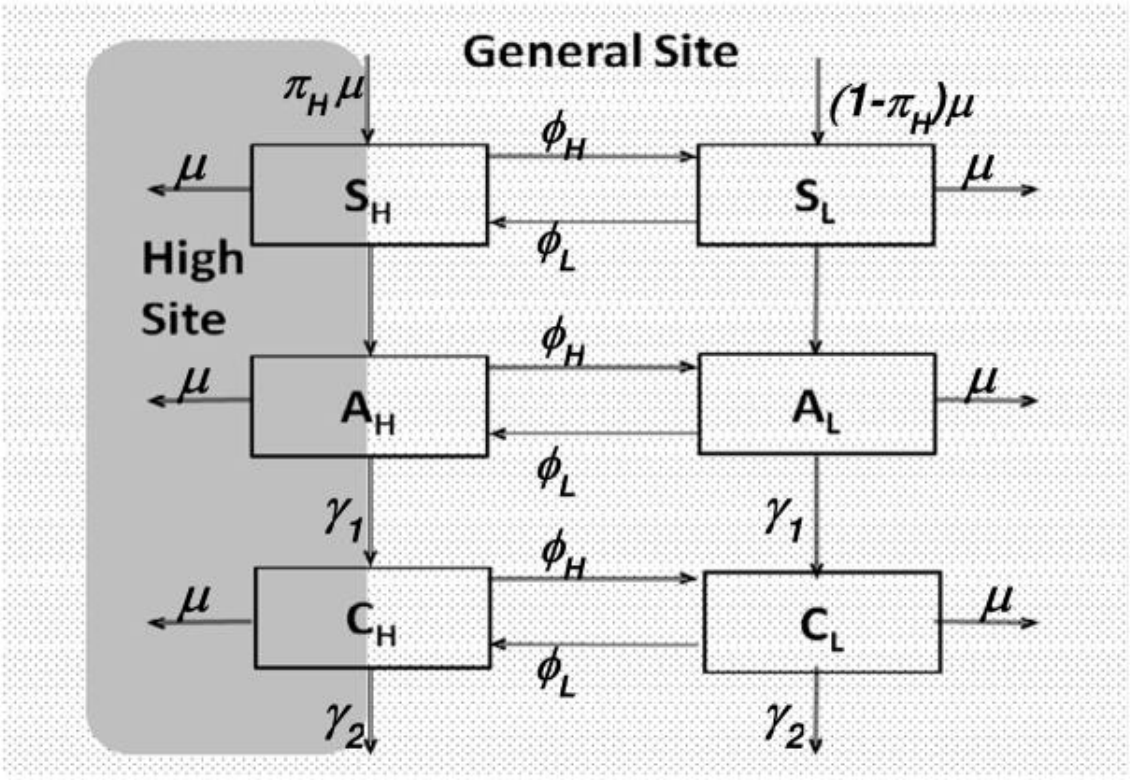


Figure 2.1: Conceptual episodic risk model with contact structure and population composition

We model episodic risk as movement between phases of high and low sexual activity. The magnitude of difference between the high and low risk phases is parameterized by the ratio of contact rates in the high and low risk phases. Average contact rates in the whole system are held constant such that an increase in this ratio or the proportion of individuals in the high-risk phase implies a decrease in the low risk contact rate to maintain a constant average contact rate. The default value for the average contact rate (χ) was selected so the system is just above the epidemic threshold under the homogenous risk setting without episodic dynamics.

The natural history of HIV infection is modeled with acute and chronic stages of infection. Contagiousness is abstracted as a higher transmissibility per act during a relatively short acute stage and a lower transmissibility during a much longer chronic stage with first order flows between stages and out of the last stage. As a simplifying assumption, we ignore the rise of contagiousness that has been demonstrated in late infection. This assumption is reasonable in a population with widespread access to antiretroviral treatments that have greatly reduced the prevalence of AIDS and the concomitant rise in viral titer associated with end-stage HIV infection. Simulations also suggested that assuming three rather than two infections stages does not fundamentally alter the phenomenon that we illustrate. In our model, we assume the average duration of the acute stage to be 2 months and that of the chronic stage to be 10 years[96]. On average, uninfected individuals will remain sexually active for 40 years.

The contagiousness parameters were adapted from an analysis of the Rakai data[131] by Pinkerton[97]. In our model, we set the baseline transmission probability (β) across the entire course of infection to be 0.003, which is consistent with overall average transmission probabilities across all homosexual sex acts (c.f. [126]). The

possible underestimate of the baseline transmission probability per partnership is made less important by the fact that we adjusted the average contact rate to be near threshold for the given baseline transmission probability.

2.3.2 Formulation of transmission probabilities and fraction of transmission potential from acute stage

The increased contagiousness in the acute stage is governed in the model by the fraction of transmission potential from the acute stage (X), which is the expected fraction of transmissions from acute HIV occurring in a homogeneous population at equilibrium. The following presents how we derived the formulation of transmission probability per contact during acute HIV infection, β_{a}, and during chronic HIV infection, β_{c}, based on baseline transmission probability, β and X.

The average time an infected individual spends in acute infection, denoted as T_1, is

$$T_1 = \frac{1}{\gamma_1 + \mu} \tag{2.1}$$

Given that the probability for an acutely infected individual to progress to the chronic stage rather than being removed due to other reasons is

$$\frac{\gamma_1}{\gamma_1 + \mu}$$

the average time an infected individual spends in the chronic stage, denoted as T_2, is

$$T_2 = \left(\frac{\gamma_1}{\gamma_1 + \mu}\right)\left(\frac{1}{\gamma_2 + \mu}\right) \tag{2.2}$$

If the rate of sexual contact in the homogeneous population is denoted as χ, the average number of secondary cases generated by an infected individual during acute infection is $\beta_1 \chi T_1$, and the average number of secondary cases generated by an infector during chronic infection is $\beta_2 \chi T_2$, while the average number of secondary cases

generated by an infector during the entire period of infection is $\beta\chi(T_1 + T_2)$. As a result, the fraction of transmission potential from acute infection is

$$X = \frac{\beta_1 \chi T_1}{\beta\chi(T_1 + T_2)} = \frac{\beta_1 T_1}{\beta(T_1 + T_2)} \tag{2.3}$$

and the fraction of transmission potential from chronic infection is

$$1 - X = \frac{\beta_2 \chi T_2}{\beta\chi(T_1 + T_2)} = \frac{\beta_2 T_2}{\beta(T_1 + T_2)} \tag{2.4}$$

After substituting equations 2.1 and 2.2 into 2.3, we find that

$$\beta_1 = X\left(1 + \frac{\gamma_1}{\gamma_2 + \mu}\right)\beta$$

Using the same logic, we find that

$$\beta_2 = (1 - X)\left(1 + \frac{\gamma_1}{\gamma_2 + \mu}\right)\beta$$

Parameterizing the system in this way, we can easily determine the effect of episodic risk on acute stage transmission rates compared to a homogeneous null model. Transmissibility per contact is assumed to be constant during each infection stage. This is determined by the above-mentioned model parameters (see Table 2.1).

2.3.3 Model elaboration for dynamic system analysis

To better clarify what flows from what stages of infection and what contact rates were sustaining transmission, we divided each of the four compartments of the infected population (A_{H},A_{L},C_{H},C_{L}) from the episodic risk model (Table 2.2) into six compartments. By doing this, we track the information of location of transmission, contact rate phase and stage of HIV infection of the source infectors. Thus, in this elaborated model, infected individuals now have a 3-letter superscript that refer to the state of the individual who infected them and where they went on to transmit. The first character in the superscripts refers to the site where they became infected,

Parameter	Value	Unit	Definition
U	$1/(40*12)$	/month	Flow of new individuals into sexually active population
μ	$1/(40*12)$	/month	Rate of leaving the sexually active population
γ_1	$1/2$	/month	Rate of transitioning from acute to chronic infection
γ_2	$1/120$	/month	Rate of death from AIDS during chronic infection
β	0.003	/contact	Average transmission probability across stages
X	variable	-	Fraction of transmission potential from acute stage
β_1	$X\beta(1+\frac{\gamma_1}{\gamma_2+\mu})$	/contact	Transmission probability during acute stage
β_2	$(1-X)\beta(1+\frac{\gamma_2+\mu}{\gamma_1})$	/contact	Transmission probability during chronic stage
π_{H}	variable	-	Average fraction of population with high contact rate in the absence of HIV
χ	3.4114	/month	Average contact rate in the entire population
ρ	variable	-	Ratio of high contact rate over low contact rate
χ_{H}	$\rho\chi_{\mathrm{L}}$	/month	Contact rate for the high-risk population
χ_{L}	$\frac{\chi}{(1-\pi_{\mathrm{H}}+\rho\pi_{\mathrm{H}})}$	/month	Contact rate for the low risk population
ν	0.3	-	Fraction of contacts of individuals with high contact rate at the high-risk site
ϕ_{H}	variable	/month	Rate of transitioning from high contact rate state to low contact rate state
ϕ_{L}	$\phi_{\mathrm{H}}(\frac{\pi_{\mathrm{H}}}{(1-\pi_{\mathrm{H}})})$	/month	Rate of transitioning from low contact rate state to high contact rate state
λ_{H}	$\frac{A_{\mathrm{H}}\beta_1+C_{\mathrm{H}}\beta_2}{S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}}}$	-	Force of infection per contact at the high risk mixing site
λ_{L}	$\frac{(A_{\mathrm{L}}+(1-\nu)A_{\mathrm{H}}\rho)\beta_1+(C_{\mathrm{L}}+(1-\nu)C_{\mathrm{H}}\rho)\beta_2}{S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}+(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}})\rho(1-\nu)}$	-	Force of infection per contact at the general mixing site

Table 2.1: List of the episodic risk model parameters and their default values

$$
\begin{aligned}
\frac{dS_{\mathrm{H}}}{dt} =& U\pi_{\mathrm{H}} - S_{\mathrm{H}}\chi_{\mathrm{H}}\left(\nu\lambda_{\mathrm{H}} + (1-\nu)\,\lambda_{\mathrm{L}}\right) - \phi_{\mathrm{H}}S_{\mathrm{H}} + \phi_{\mathrm{L}}S_{\mathrm{L}} - \mu S_{\mathrm{H}} \\
\frac{dS_{\mathrm{L}}}{dt} =& U\left(1-\pi_{\mathrm{H}}\right) - S_{\mathrm{L}}\chi_{\mathrm{L}}\lambda_{\mathrm{L}} + \phi_{\mathrm{H}}S_{\mathrm{H}} - \phi_{\mathrm{L}}S_{\mathrm{L}} - \mu S_{\mathrm{L}} \\
\frac{dA_{\mathrm{H}}}{dt} =& S_{\mathrm{H}}\chi_{\mathrm{H}}\left(\nu\lambda_{\mathrm{H}} + (1-\nu)\,\lambda_{\mathrm{L}}\right) - A_{\mathrm{H}}\gamma_1 - \phi_{\mathrm{H}}A_{\mathrm{H}} + \phi_{\mathrm{L}}A_{\mathrm{L}} - \mu A_{\mathrm{H}} \\
\frac{dC_{\mathrm{H}}}{dt} =& A_{\mathrm{H}}\gamma_1 - C_{\mathrm{H}}\gamma_2 - \phi_{\mathrm{H}}C_{\mathrm{H}} + \phi_{\mathrm{L}}C_{\mathrm{L}} - \mu C_{\mathrm{H}} \\
\frac{dA_{\mathrm{L}}}{dt} =& S_{\mathrm{L}}\chi_{\mathrm{L}}\lambda_{\mathrm{L}} - A_{\mathrm{L}}\gamma_1 - \phi_{\mathrm{L}}A_{\mathrm{L}} + \phi_{\mathrm{H}}A_{\mathrm{H}} - \mu A_{\mathrm{L}} \\
\frac{dC_{\mathrm{L}}}{dt} =& A_{\mathrm{L}}\gamma_1 - C_{\mathrm{L}}\gamma_2 - \phi_{\mathrm{L}}C_{\mathrm{L}} + \phi_{\mathrm{H}}C_{\mathrm{H}} - \mu C_{\mathrm{L}}
\end{aligned}
$$

Table 2.2: Differential equations for the deterministic episodic risk compartmental model

with R being the high-risk mixing site and G being the common mixing site. The middle letter refer to the contact rate phase of the source infector, with H being the high contact rate phase and L being the low contact rate phase. The last character represents the stage of HIV infection of the source infector, with A being the acute stage and C being the chronic stage. The set of differential equations for the dynamic system analysis of the elaborated model is provided in Tables 2.3 and 2.4.

2.4 Results

2.4.1 Exploring joint effects of high-risk turnover rate and between–risk-group contact rate ratio at the population-level

We numerically solved equations out to equilibrium. The rate of turnover in the high-risk group had strong effects on both the endemic prevalence and the fraction of transmission during acute stage. As the turnover rate of the high-risk group slows from 10 times per month (once every 0.008 years) to once per 1000 months (once every 83.3 years), both the endemic prevalence and the acute transmission rate show an inverted U shape, peaking between turnover rates of 1 per month and 1 per year (Figure 2.2). At the extremes of fast or slow turnover, the fraction of acute transmissions is the same as in a homogeneous population. As the ratios of contact

$$
\begin{aligned}
\frac{dA_{\mathrm{H}}^{\mathrm{RHA}}}{dt} =& S_{\mathrm{H}}\chi_{\mathrm{H}}\nu\frac{A_{\mathrm{H}}\beta_1}{S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}}} \\
&- A_{\mathrm{H}}^{\mathrm{RHA}}\gamma_1 - \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{RHA}} + \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{RHA}} - \mu A_{\mathrm{H}}^{\mathrm{RHA}} \\
\frac{dA_{\mathrm{H}}^{\mathrm{RHC}}}{dt} =& S_{\mathrm{H}}\chi_{\mathrm{H}}\nu\frac{C_{\mathrm{H}}\beta_2}{S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}}} \\
&- A_{\mathrm{H}}^{\mathrm{RHC}}\gamma_1 - \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{RHC}} + \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{RHC}} - \mu A_{\mathrm{H}}^{\mathrm{RHC}} \\
\frac{dA_{\mathrm{H}}^{\mathrm{GHA}}}{dt} =& S_{\mathrm{H}}\chi_{\mathrm{H}}\left(1-\nu\right)\frac{A_{\mathrm{H}}\left(1-\nu\right)\rho\beta_1}{\left(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}}\right)\left(1-\nu\right)\rho+S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
&- A_{\mathrm{H}}^{\mathrm{GHA}}\gamma_1 - \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GHA}} + \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GHA}} - \mu A_{\mathrm{H}}^{\mathrm{GHA}} \\
\frac{dA_{\mathrm{H}}^{\mathrm{GHC}}}{dt} =& S_{\mathrm{H}}\chi_{\mathrm{H}}\left(1-\nu\right)\frac{C_{\mathrm{H}}\left(1-\nu\right)\rho\beta_2}{\left(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}}\right)\left(1-\nu\right)\rho+S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
&- A_{\mathrm{H}}^{\mathrm{GHC}}\gamma_1 - \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GHC}} + \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GHC}} - \mu A_{\mathrm{H}}^{\mathrm{GHC}} \\
\frac{dA_{\mathrm{H}}^{\mathrm{GLA}}}{dt} =& S_{\mathrm{H}}\chi_{\mathrm{H}}\left(1-\nu\right)\frac{A_{\mathrm{L}}\beta_1}{\left(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}}\right)\left(1-\nu\right)\rho+S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
&- A_{\mathrm{H}}^{\mathrm{GLA}}\gamma_1 - \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GLA}} + \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GLA}} - \mu A_{\mathrm{H}}^{\mathrm{GLA}} \\
\frac{dA_{\mathrm{H}}^{\mathrm{GLC}}}{dt} =& S_{\mathrm{H}}\chi_{\mathrm{H}}\left(1-\nu\right)\frac{C_{\mathrm{L}}\beta_2}{\left(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}}\right)\left(1-\nu\right)\rho+S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
&- A_{\mathrm{H}}^{\mathrm{GLC}}\gamma_1 - \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GLC}} + \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GLC}} - \mu A_{\mathrm{H}}^{\mathrm{GLC}} \\
\frac{dC_{\mathrm{H}}^{\mathrm{RHA}}}{dt} =& A_{\mathrm{H}}^{\mathrm{RHA}}\gamma_1 - C_{\mathrm{H}}^{\mathrm{RHA}}\gamma_2 - \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{RHA}} + \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{RHA}} - \mu C_{\mathrm{H}}^{\mathrm{RHA}} \\
\frac{dC_{\mathrm{H}}^{\mathrm{RHC}}}{dt} =& A_{\mathrm{H}}^{\mathrm{RHC}}\gamma_1 - C_{\mathrm{H}}^{\mathrm{RHC}}\gamma_2 - \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{RHC}} + \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{RHC}} - \mu C_{\mathrm{H}}^{\mathrm{RHC}} \\
\frac{dC_{\mathrm{H}}^{\mathrm{GHA}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GHA}}\gamma_1 - C_{\mathrm{H}}^{\mathrm{GHA}}\gamma_2 - \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GHA}} + \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GHA}} - \mu C_{\mathrm{H}}^{\mathrm{GHA}} \\
\frac{dC_{\mathrm{H}}^{\mathrm{GHC}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GHC}}\gamma_1 - C_{\mathrm{H}}^{\mathrm{GHC}}\gamma_2 - \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GHC}} + \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GHC}} - \mu C_{\mathrm{H}}^{\mathrm{GHC}} \\
\frac{dC_{\mathrm{H}}^{\mathrm{GLA}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GLA}}\gamma_1 - C_{\mathrm{H}}^{\mathrm{GLA}}\gamma_2 - \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GLA}} + \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GLA}} - \mu C_{\mathrm{H}}^{\mathrm{GLA}} \\
\frac{dC_{\mathrm{H}}^{\mathrm{GLC}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GLC}}\gamma_1 - C_{\mathrm{H}}^{\mathrm{GLC}}\gamma_2 - \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GLC}} + \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GLC}} - \mu C_{\mathrm{H}}^{\mathrm{GLC}}
\end{aligned}
$$

Table 2.3: Part 1 of the equations for the elaborated deterministic compartmental model focusing on the source and site of infection: differential equations for the size of high-risk infected subpopulations. The corresponding equations for low-risk infected subpopulations are contained in Table 2.4.

$$
\begin{aligned}
\frac{dA_{\mathrm{L}}^{\mathrm{RHA}}}{dt} =& -A_{\mathrm{L}}^{\mathrm{RHA}}\gamma_1 + \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{RHA}} - \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{RHA}} - \mu A_{\mathrm{L}}^{\mathrm{RHA}} \\
\frac{dA_{\mathrm{L}}^{\mathrm{RHC}}}{dt} =& -A_{\mathrm{L}}^{\mathrm{RHC}}\gamma_1 + \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{RHC}} - \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{RHC}} - \mu A_{\mathrm{L}}^{\mathrm{RHC}} \\
\frac{dA_{\mathrm{L}}^{\mathrm{GHA}}}{dt} =& S_{\mathrm{L}}\chi_{\mathrm{L}}\frac{A_{\mathrm{H}}\beta_1(1-\nu)\rho}{(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}})(1-\nu)\rho + S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
& - A_{\mathrm{L}}^{\mathrm{GHA}}\gamma_1 + \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GHA}} - \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GHA}} - \mu A_{\mathrm{L}}^{\mathrm{GHA}} \\
\frac{dA_{\mathrm{L}}^{\mathrm{GHC}}}{dt} =& S_{\mathrm{L}}\chi_{\mathrm{L}}\frac{C_{\mathrm{H}}\beta_2(1-\nu)\rho}{(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}})(1-\nu)\rho + S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
& - A_{\mathrm{L}}^{\mathrm{GHC}}\gamma_1 + \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GHC}} - \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GHC}} - \mu A_{\mathrm{L}}^{\mathrm{GHC}} \\
\frac{dA_{\mathrm{L}}^{\mathrm{GLA}}}{dt} =& S_{\mathrm{L}}\chi_{\mathrm{L}}\frac{A_{\mathrm{L}}\beta_1}{(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}})(1-\nu)\rho + S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
& - A_{\mathrm{L}}^{\mathrm{GLA}}\gamma_1 + \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GLA}} - \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GLA}} - \mu A_{\mathrm{L}}^{\mathrm{GLA}} \\
\frac{dA_{\mathrm{L}}^{\mathrm{GLC}}}{dt} =& S_{\mathrm{L}}\chi_{\mathrm{L}}\frac{C_{\mathrm{L}}\beta_2}{(S_{\mathrm{H}}+A_{\mathrm{H}}+C_{\mathrm{H}})(1-\nu)\rho + S_{\mathrm{L}}+A_{\mathrm{L}}+C_{\mathrm{L}}} \\
& - A_{\mathrm{L}}^{\mathrm{GLC}}\gamma_1 + \phi_{\mathrm{H}}A_{\mathrm{H}}^{\mathrm{GLC}} - \phi_{\mathrm{L}}A_{\mathrm{L}}^{\mathrm{GLC}} - \mu A_{\mathrm{L}}^{\mathrm{GLC}} \\
\frac{dC_{\mathrm{L}}^{\mathrm{RHA}}}{dt} =& A_{\mathrm{H}}^{\mathrm{RHA}}\gamma_1 - C_{\mathrm{L}}^{\mathrm{RHA}}\gamma_2 - \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{RHA}} + \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{RHA}} - \mu C_{\mathrm{L}}^{\mathrm{RHA}} \\
\frac{dC_{\mathrm{L}}^{\mathrm{RHC}}}{dt} =& A_{\mathrm{H}}^{\mathrm{RHC}}\gamma_1 - C_{\mathrm{L}}^{\mathrm{RHC}}\gamma_2 - \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{RHC}} + \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{RHC}} - \mu C_{\mathrm{L}}^{\mathrm{RHC}} \\
\frac{dC_{\mathrm{L}}^{\mathrm{GHA}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GHA}}\gamma_1 - C_{\mathrm{L}}^{\mathrm{GHA}}\gamma_2 - \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GHA}} + \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GHA}} - \mu C_{\mathrm{L}}^{\mathrm{GHA}} \\
\frac{dC_{\mathrm{L}}^{\mathrm{GHC}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GHC}}\gamma_1 - C_{\mathrm{L}}^{\mathrm{GHC}}\gamma_2 - \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GHC}} + \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GHC}} - \mu C_{\mathrm{L}}^{\mathrm{GHC}} \\
\frac{dC_{\mathrm{L}}^{\mathrm{GLA}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GLA}}\gamma_1 - C_{\mathrm{L}}^{\mathrm{GLA}}\gamma_2 - \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GLA}} + \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GLA}} - \mu C_{\mathrm{L}}^{\mathrm{GLA}} \\
\frac{dC_{\mathrm{L}}^{\mathrm{GLC}}}{dt} =& A_{\mathrm{H}}^{\mathrm{GLC}}\gamma_1 - C_{\mathrm{L}}^{\mathrm{GLC}}\gamma_2 - \phi_{\mathrm{L}}C_{\mathrm{L}}^{\mathrm{GLC}} + \phi_{\mathrm{H}}C_{\mathrm{H}}^{\mathrm{GLC}} - \mu C_{\mathrm{L}}^{\mathrm{GLC}}
\end{aligned}
$$

Table 2.4: Part 2 of the equations for the elaborated deterministic compartmental model focusing on the source and site of infection: differential equations for the size of low-risk infected subpopulations. The corresponding equations for high-risk infected subpopulations are contained in Table 2.3.

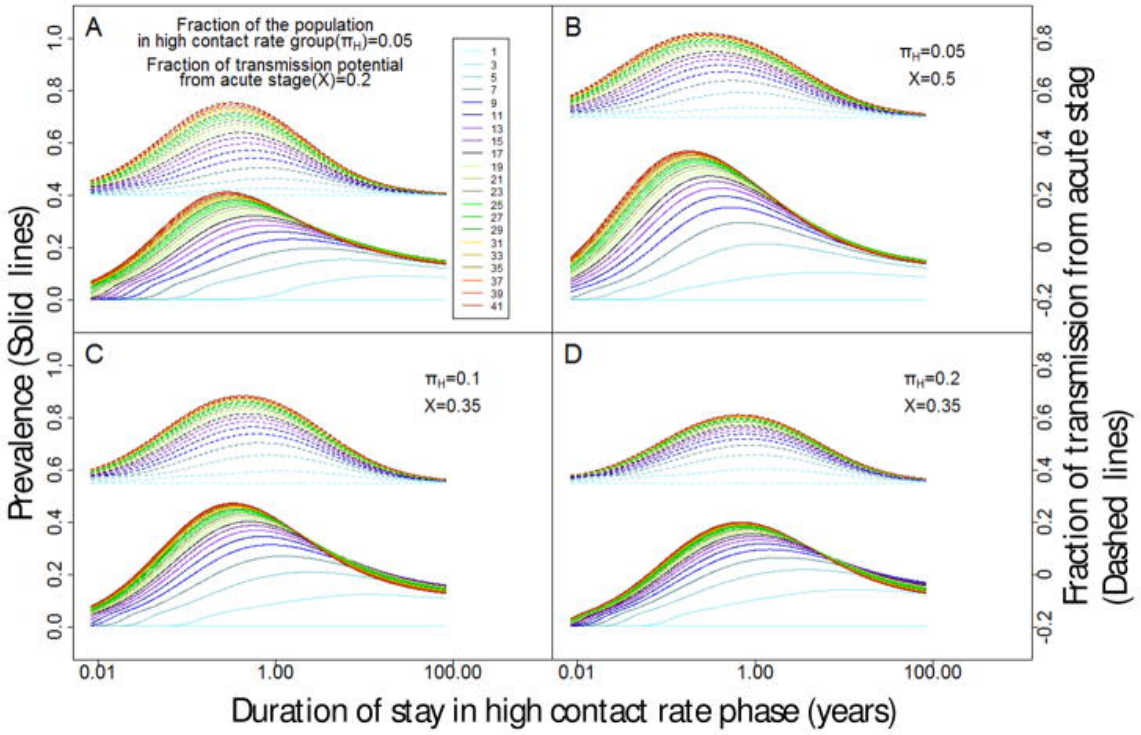


Figure 2.2: Endemic prevalence and fraction of transmission during acute stage of infection as a function of duration of stay at high risk level (on a log scale) as the contact rate ratio (r) is raised from 1 to 41 by increments of 2 and shown in different colors. The left y axis is for the endemic prevalence and the right y axis is for the fraction of transmission during the acute stage of infection. Parameters values not specified in this figure are consistent with those in Table 2.1.

rates between high and low risk groups increase, the peak effects of turnover rates increase as well and occur at faster turnover rates.

Increasing the fraction of transmission potential during acute stage while keeping the total transmission potential constant increases both the endemic fraction of transmissions during acute stage and the endemic prevalence (compare panels A and B). Furthermore, the endemic prevalence peaks at a faster turnover rate. This is because, given an increased transmission rate during the acute stage, individuals need to spend shorter time at high-risk phase to lift the endemic prevalence to peak. These observed effects hold at all values of the contact rate ratio as shown in Figure 2.2.

As the average fraction of time that an individual spends at high risk (π_H) increases, both the endemic fraction of transmission during the acute stage and the endemic prevalence decrease at each value of the contact rate ratio (compare panels C and D in Figure 2.2). Given our assumption of constant average contact rate, for any given contact rate ratio greater than 1, increasing π_H causes a decrease in contact rates for both risk levels and hence decreases the overall force of infection. When the turnover rate is low enough and the contact rate ratio is high enough, the endemic prevalence starts to decrease as the contact rate ratio increases. Given a slow turnover, high-risk susceptible individuals are infected much faster as compared to the low risk susceptible individuals due to their elevated contact rates causing infection saturation in the high-risk phase. Higher turnover rates supplement the pool of high-risk susceptible individuals with new high-risk susceptible individuals. As π_H increases, each individual spends more time in the high-risk phase. This increases the proportion of contacts between infected individuals causing saturation of infection in the high-risk phase.

2.4.2 Effect of separate mixing on the endemic prevalence and fraction of transmission by acute stage

Overall, the fraction of contacts made by high-risk infectors at the high-risk site (ν) showed little impact either on the endemic prevalence or on fraction of transmission during acute stage or on the dynamics of these two outcomes (data not shown). This is unlike the observations in the 2005 paper by Koopman et al.[79]. Only at lower overall contact rates and very slow turnover between contact rate groups do we see the effects of ν as noted in the 2005 Koopman et al. paper[79]. At any other contact rates, a higher fraction of high contact rate group's contacts made at a high risk mixing site increases the rate of exponential growth in the early epidemic.

2.4.3 Exploration of dynamics underlying high contact rate group turnover effect

We now look into the dynamics of prevalence and the fraction of transmissions from acute HIV infection (AHI) over a period of 150 years. In Figure 2.3 (A and B), we explore these dynamics for the duration of time spent in the high-risk state for ranges from 0.01 to 80 years. In Figure 2.3 (C and D), we explore the dynamics by varying the ratio of contact rates from 1 to 41.

Figure 2.3 (A and B) shows that episodic risk increases the contribution of acute HIV early in the epidemic and later at the endemic equilibrium. Faster risk turnover results in the faster replenishment of the susceptible population in the high-risk phase. At the same time, it increases the risk of a low-risk susceptible contacting an infected individual who is still in the highly contagious acute stage. For slower turnover rates, the fraction of the transmissions from AHI decreases. This is because of the decrease in the inflow of susceptibles into the high-risk phase. Slower turnover also increases the duration of the time spent in the high-risk phase by high-risk individuals. Thus, high-risk individuals are likely to progress to the chronic stage by the time they switch to the low-risk phase.

Overall, the prevalence and the fraction of transmissions during the acute stage increase with the ratio of contact rates (see Figure 2.3; panels C and D). Increasing contact rate ratios result in increasing high-risk contact rates that in turn increase transmission from high-risk infectors. For high contact rate ratios, the fraction of transmission from acute stage reaches a higher peak level, indicating a greater effect when the high-risk contact rate increases. Higher peak levels of acute stage transmissions reflect the potential of the acute HIV population to build the epidemic together with the replenishment of the susceptible population. This explains why the fraction of transmissions during acute stage with a higher peak level also has a

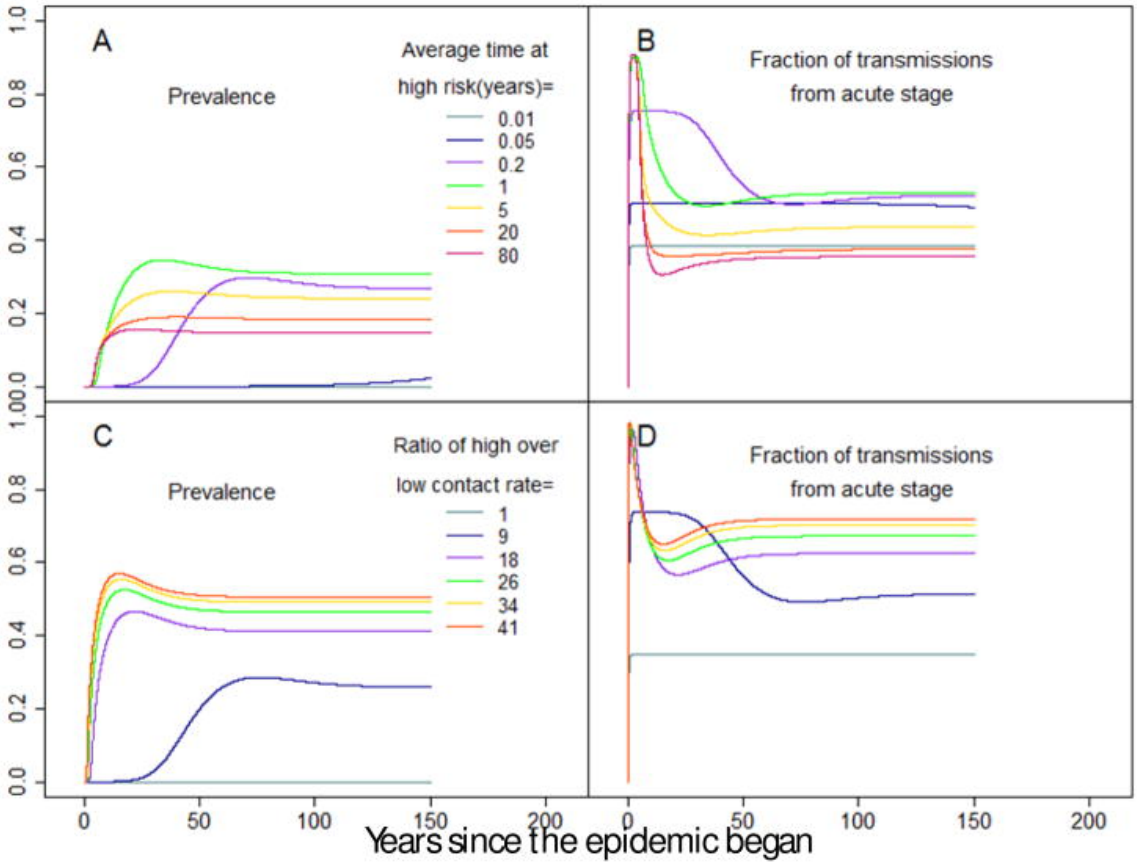


Figure 2.3: Model runs showing the endemic prevalence (Panels A and C) and the fraction of transmissions from the acute stage (Panels B and D) for 150 years. Above: average time spent at high contact rate phase is raised from 0.1 to 80 years. Below: the ratio of high over low contact rate from 1 to 41. Parameter values not specified in this figure are the same as given in Table 2.1.

larger rebound as shown in Figure 2.3 (B and D).

2.4.4 Effect of risk turnover and contact rate ratios on the contact rate of acute HIV stage population

In Figure 2.4, we examine the effects of high-risk group turnover rate on the endemic prevalence, the average contact rates of the acute stage population and that of the susceptibles at equilibrium. We explore this over a range of the duration of time spent in the high-risk phase from 0.008 years to 83 years and a range of the ratio of contact rates from 1 to 41. As Figure 2.4B shows, the turnover rate of the high-risk population has a clear effect on the average contact rates of acute stage

infected individuals at equilibrium. Higher turnover rate brings more susceptibles into the high contact rate phase. At higher contact rate ratios, the new susceptible at high contact rate are more likely to be infected and thus faster turnover in the high contact rate phase has even greater effects.

The endemic prevalence reaches its peak at a faster turnover rate than the average contact rate of acutely infected individuals (Figure 2.4A). This is due to the combined effect of the overall force of infection and average contact rate among the susceptibles (Figure 2.4B). Given the high contagiousness of the acute stage infectors, the interaction of high transmissibility and increased contact rates during acute stage drives the overall force of infection.

2.4.5 Tracking sources of infections

By breaking the compartments of infections in our original model into different subgroups, we were able to track the source of infection and mixing site where the transmissions occurred. Figure 2.5A and 2.5C illustrate the fraction of transmissions during acute stage and chronic stage HIV infection. Figures 2.5B and 2.5D focus on the contributions of high-risk infectors and low risk infectors. From B and D, we can see that, although the high contact rate group only comprises of 5% of the total population, they account for 50% or more of all transmissions at contact rate ratios as low as 7. Although there is an effect of the turnover rate on the fraction of infections by individuals in the high-risk phase, the effect is small. The role of the high versus low-risk infectors is controlled primarily by the contact rate ratio.

From panels A and C in Figure 2.5, we see a different role for the contact rate ratio and the turnover rate in determining the fraction of transmission from acute stage infectors. The contact rate ratio has little effect on the fraction of infections from acute infectors at extreme values of the turnover rate. However, at intermediate

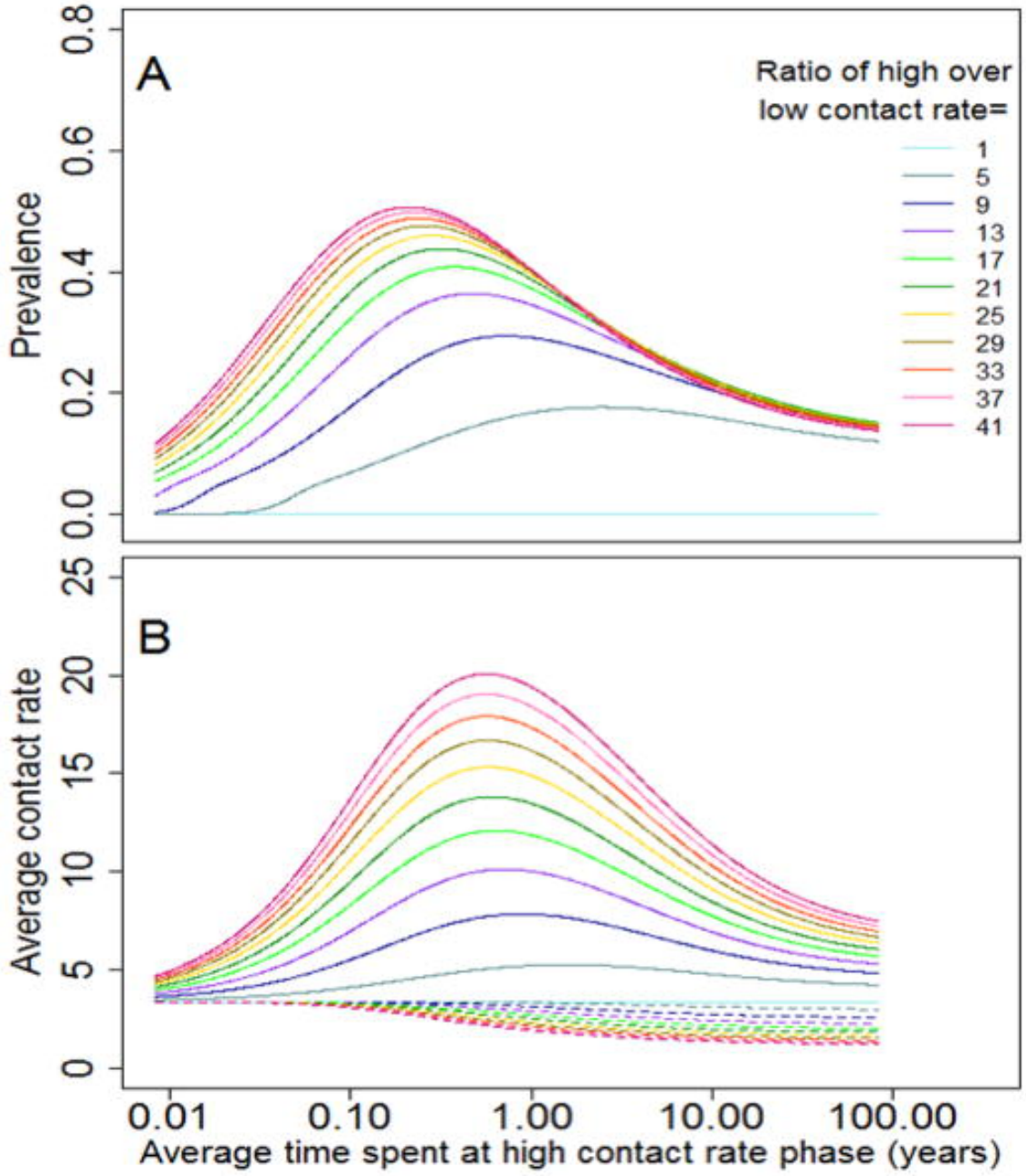


Figure 2.4: Endemic prevalence (A) and average contact rate of acutely infected individuals (solid lines in B) or average contact rate of susceptibles (dashed lines in B) at equilibrium as average time spent at high contact rate phase is varied from 0.01 to 100 years. Parameter values not specified in this figure are the same as given in Table 2.1.

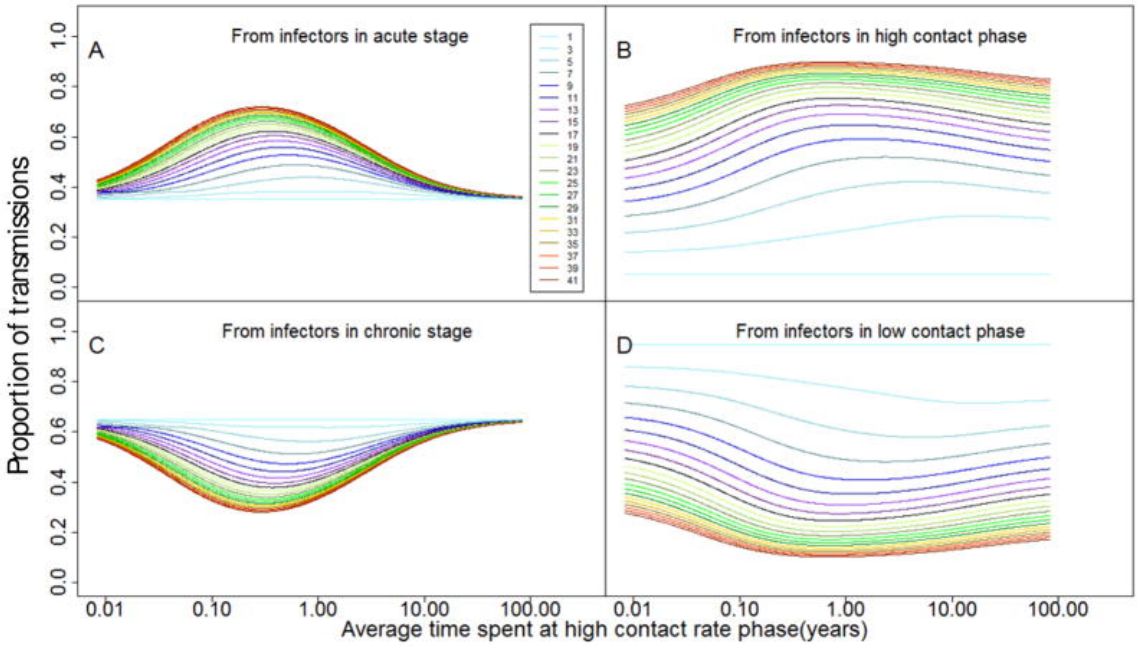


Figure 2.5: Proportion of transmissions at endemic equilibrium generated by acutely infected individuals(A), infected individuals in high contact rate phase (B), chronically infected individuals (C) or infected individuals in low contact rate phase (D). In each panel, the ratio of high over low contact rate is raised from 1 to 41 by increments of 2. Parameter values not specified in this figure are the same as given in Table 2.1.

values of the turnover rate, the contact rates ratio can nearly double the contribution of acute stage infectors. This interaction is caused by the increased prevalence of high-risk, acutely infected individuals. The flow of new susceptibles into the high-risk phase generates an increased incidence and therefore an increased prevalence of acute stage, high-risk infectors. Because the dynamics are assumed to be mass action, the increased prevalence of acute high-risk infectors means increased contact between susceptibles and acutely infected individuals. The effect diminishes at extremely short average durations of the high risk phase because the system becomes essentially homogeneous. In that case, the system is controlled by the average contact rate rather than the contact rate ratio.

At extremely long average durations, on the other hand, any newly infected individual is highly unlikely to change risk groups prior to leaving the system. He will, therefore, with high probability, have the same contact rate for the duration of his infection, and the fraction of his expected cumulative contribution to the force of infection over time that occurs during his acute infection will be essentially the same as in a homogeneous system. Because this is true for all infected individuals, it will also be true for the system as a whole. In this limit, the contact rate ratio still matters for determining the total incidence, but does not affect the fraction of that incidence that comes from transmissions by acutely or chronically infected individuals.

In Figure 2.6, we further divide the source of infection based on their contact rate phase, stage of infection and location. When the turnover rate is low, neither acute nor chronic individuals are generating many new infections at the high-risk site (panels A and D). This is due to the fact that without turnover to replenish susceptibles in the high-risk phase, the high-risk individuals rapidly become infected. Therefore, at the high-risk site almost all contacts are among already infected individuals. How-

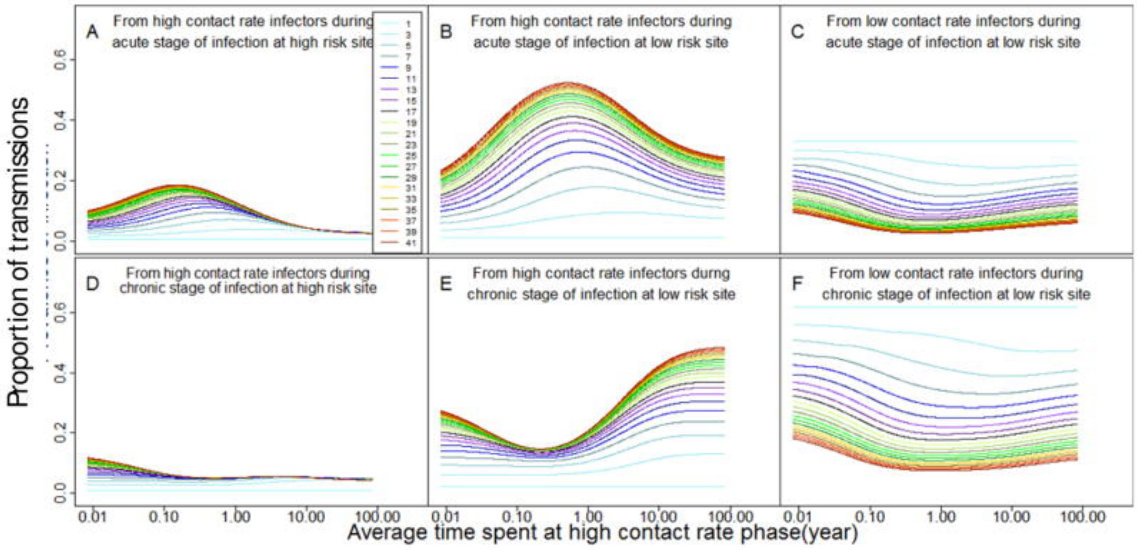


Figure 2.6: Proportion of transmissions at endemic equilibrium generated by different types of infected individuals. In each panel, the ratio of high over low contact rate is raised from 1 to 41 by increments of 2. Parameter values not specified in this figure are the same as given in Table 2.1.

ever, at the low-risk site, susceptibles with low contact rates are much more likely to contact either an acute (panel B) or a chronic (panel E) high risk infected individual resulting in elevated incidence rates from those classes in the absence of turnover. Also, by assumption, most individuals entering the model are susceptibles with low contact rates, which provides a greater supply of susceptibles to the low risk mixing site.

2.5 Discussion

2.5.1 Explaining the major effects we observed

We have shown that episodic periods of high-risk behavior interspersed with longer periods of lower risk have large effects on the endemic prevalence of HIV infection and the fraction of transmissions from acute HIV infection. We see that prevalence can vary from near zero to very high levels and the fraction of transmissions from acute infection can nearly double as the turnover rate from low-to-high contact rate

phases increase and as the time spent with high rates varies. This is despite the fact that the total contacts in the population and the transmission potential for each individual are kept constant during each stage of infection.

Two important elements explain the strong effects mentioned above. The first has to do with the average total contagiousness of individuals during acute and chronic infection. That contagiousness is determined jointly by the number of contacts made and transmission probabilities per contact. Episodic risk increases the association between high contact rates and the high transmissibility of acute infection because an individual is more likely to get infected during a high contact rate period and is more likely to still be in that high contact rate period during their acute HIV infection (AHI). Then, as they progress to the less contagious chronic infection stage, they also pass into a lower contact rate phase. The more strongly that the high transmission stage of infection is correlated with the high contact rate state, the higher will be the average contagiousness across the course of infection. This is reflected in the part of Figure 2.4 that presents the average contact rates during the acute infection stage.

The second part of the explanation for these big effects has to do with the average contact rates of susceptible individuals. When individuals spend a long time in the high contact rate state, they become infected and a sexual contact by them with other infected individuals no longer presents an opportunity to transmit infection. If individuals stay in a low contact rate state, they are more likely to avoid infection. Then when they transit to a high contact rate state, they provide a new source of high contact rate susceptible individuals that the individuals with high contact rates in the acute infection stage can infect. This is seen in the contact rates of susceptible individuals in Figure 2.4. The average contact rate of susceptible individuals is seen to increase considerably as the turnover increases. When the turnover rate is slow,

high contact rate individuals are infected and are not replaced by new susceptible individuals. This lowers the average contact rate of susceptible individuals.

Our previous work on when to focus control on high risk groups[79] led us to believe that a third element would also act to generate episodic risk effects on prevalence and AHI transmissions. That previous work modeled SIR infections. For SIR infections, we saw that when high-risk individuals mixed separately from low risk individuals, it allowed them to build up higher levels of infection so that they then disseminated more infection when they made contact with lower risk individuals. However, that was not the case for the HIV infections modeled here. Whether high contact individuals mixed at a high-risk or a common mixing site, the prevalence and the fraction of transmissions from acute infection were barely changed at all. The difference was not that in the previous work we allowed transmission probabilities rather than contact rates to define high risk. In fact, we saw the same effects of separate high risk mixing in that previous work when we defined high risks by contact. In further support of the cause of high risk (contact rates vs. transmission probabilities), we have observed that mixing makes little difference when high risk is caused by higher transmission probabilities (data not presented). The reasons for this are complex: under some conditions, separate mixing at a high-risk site can lead to a greater dissemination of infection. However, discussion of this effect is beyond the scope of this chapter.

2.5.2 The real-world importance of the described theoretical effects

If the episodic risk effects that we have described occur in the real world to the extent that we suspect they do, then previous model analyses using fixed behaviors for subgroups[22, 104] have underestimated the fraction of transmissions from acute HIV. Nevertheless, a key question is how we can assess the extent to which these

episodic risk effects are occurring in any particular population. Fluctuation of contact rates was observed in the CDC cohort studied to assess HIV transmission risks from different acts[108]. That study included diverse geographic areas. The existence of episodic risk behavior is less formally supported by experiences of HIV counselors where most people with high-risk behaviors report that they have had longer periods of low risk behaviors.

Our results may be compared to recent modeling studies that have focused on the role of acute infection[37, 100]. Eaton et al.[37] analyze an important phenomenon that, like episodic risk, augments the importance of acute infection. In some sense, entering a high contact rate phase may be akin to entering a concurrent partnership phase, although in our model, we assume instantaneous contacts. On the other hand, the study by Powers et al.[100] assume fixed risk, which corresponds to our results when the duration of stay in high risk is very long such that there is virtually no risk turnover. The Bayesian melding approach to model fitting used by Powers et al.[100] means that the effects of the phenomenon of episodic risk could be absorbed partially into the transmission by stage parameter estimates. It seems that their results would have estimated a larger fraction of transmissions from acute infection if they had included episodic risk in their model. However, this is hard to assess. Broader evidence should be sought regarding whether there are real world effects of episodic risk corresponding to the theoretical effects we have demonstrated. One approach could be to do more behavioral studies that ask the needed questions. So far, few studies have documented episodic risk behavior. Another approach is to find signatures of episodic risk in both HIV surveillance data and in HIV genetic sequences[19]. When the model in this chapter is simulated as a stochastic discrete individual-based model, we observe that infection trees of acute HIV transmissions

are interspersed among more isolated chronic infection transmissions and have size distributions that are roughly comparable to those observed in Montreal[18]. Our initial results show that episodic risk dynamics influence both the size and duration of acute HIV outbreaks providing a possible link between genetic cluster size distributions and episodic risk dynamics. More sophisticated methods that are currently being developed can be used to calculate the likelihood of the genetic data given a transmission model allowing for direct evaluation of the evidence for episodic risk in the genetic data[105, 127, 129]. Parameter ranges should be further constricted by fitting models to observed patterns of HIV infection over time using surveillance data.

More thorough model explorations may be needed to assess the real world importance of episodic risk effects. More surprises regarding dynamic effects may arise as we analyze the interactions of episodic risk effects with contact pattern effects, partnership duration, insertive-receptive behavior effects, and oral, anal, protected, and unprotected risk pattern effects. To increase the robustness of our inferences (that episodic risk behavior greatly increases HIV prevalence and acute HIV transmissions), a thorough exploration of the effects of relaxing the simplifying assumptions in the model presented is called for. For example, realistic distributions for the duration of acute stage and the distribution of virus levels across that time should be explored. In reality, there is a high narrow peak of virus levels a few weeks into infection and then a long period of falling levels before the set point of chronic infection is reached[95]. We hypothesize that the effects we have observed will be just as strong or stronger under more realistically detailed acute infection timing and contagiousness. Likewise, our initial explorations indicate that a more realistic description of episodic risk behavior in a population will not significantly diminish the effects we

have illustrated in this chapter.

The infection control implications of our findings suggest that some HIV transmission control activities may be misdirected. For example, the test and treat strategy by default detects mostly cases after acute infection. Our results indicate that greater focusing on acute infection transmission clusters[19] will be needed to detect more acute infections. Second, behavior modification messages may be similarly misdirected. In our model, we did not incorporate change in sexual behavior over an individual's lifetime because of, e.g. HIV awareness programs. This is because our aim in this chapter is to understand the dynamics of episodic risk behavior under theoretical settings. Nevertheless, if suggestions to decrease risk only work for a period but are overwhelmed by situational pressures, then our model indicates that such suggestions could not only fail to decrease transmission, but might in fact increase transmission.

2.5.3 The broader context of this work and future directions

The analyses presented in this chapter are a needed step toward improving our understanding of HIV transmission dynamics. All perception comes from a combination of data and theory. We do not see patterns when our minds are not prepared to see them. To an extent our ignorance of HIV transmission patterns in MSM is attributable to a dearth of well-structured theory about how different hypothetical transmission system conformations lead to specific transmission patterns and what mechanisms underlie these effects. A body of HIV transmission system theory is growing and thousands of modeling studies of HIV transmission have been published[3]. Still, theory and data are barely beginning to come together in a way that allows us to see clearly how and why acute HIV transmissions cluster, and how this could explain the failure to control the spread of HIV infection in the MSM

population.

By relaxing some of the unrealistic assumptions in our model and incorporating realistic details such as oral and anal acts, insertive and receptive behavior, partnership duration, and mixing patterns, we will begin to identify characteristics of both surveillance data and genetic sequence data that indicate what is affecting transmission dynamics. This, we believe, will clarify the role acute infection transmissions plays in the MSM HIV transmission system.

CHAPTER III

Strong Influence of Behavioral Dynamics on the Ability of Testing and Treating HIV to Stop Transmission

3.1 Chapter preface

The contents of this chapter and Appendix A are adapted from the main text and Supplementary Information, respectively, of a paper[58] originally published in Scientific Reports.

In this chapter, I extend the model of HIV transmission dynamics presented in Chapter II to incorporate (a simplified version of) universal test and treat (UT&T), a strategy for control of the HIV epidemic that relies on encouraging the routine, voluntary HIV testing of at-risk individuals, with a goal of immediately starting antiretroviral therapy (ART) for those who test positive, in order to achieve viral suppression and thereby reduce or (effectively) eliminate their ability to generate secondary transmissions[36]. I also extend previous algebraic results regarding the effect of static risk heterogeneity on the basic reproduction number (R_0) to account for the effects of risk heterogeneity being episodic rather than static. By applying these results to the extended model, I obtain the counterintuitive result that episodic risk can in many cases make control of transmission through universal test and treat easier, despite increasing the fraction of transmissions that come from early infection.

There are a few minor changes in the presentation of the model in this chapter,

relative to Chapter II, apart from the addition of treatment. The largest is the replacement of a parameterization of episodic risk in terms of the rate of *transition* between risk groups with a parameterization in terms of the rate of random *reselection* of what risk group an individual is in, allowing for the possibility that the old and new groups will be the same. The dynamics of the model are not altered, only how those dynamics are expressed. One major advantage of this change in parameterization is that the rate of reselection is equal for the high- and low-risk groups, unlike the rate of transition. Because of this, the model as presented in this chapter has a natural and straightforward generalization to a model including three or more risk groups (or even a continuous distribution of contact rates), which the model as presented in Chapter II does not.

There are also a few changes in the symbols used to denote certain variables, without changes in the meaning of those variables themselves.

3.2 Introduction

The ultimate goal of Treatment as Prevention (TasP), the complete elimination of HIV transmission, will be achieved if the number of transmissions made by a "typical" infected individual over the course of his infection can be reduced to less than one[25]. There are three major factors that affect how difficult it is to achieve this goal through Universal Test and Treat (UT&T): (1) The series of steps that must be traversed from diagnosis to viral suppression (the treatment cascade), (2) the fraction of transmissions that occur early in the transmitter's infected period (which is determined by both biological and social factors), and (3) the basic reproduction number of the virus within the population of interest (which is also affected by both biological and social factors). The basic reproduction number (R_0) and the fraction

of early transmissions relate to two dimensions of difficulty in achieving elimination through UT&T: The basic reproduction number is closely related to the fraction of transmissions that must be prevented through diagnosis and treatment in order to achieve elimination (which is $1 - \frac{1}{R_0}$), and early transmissions are more difficult to prevent with UT&T than later ones.

The importance of the treatment cascade and the fraction of transmissions from early infection have been extensively discussed, but the importance of the basic reproduction number (R_0), although touched on in some articles[21, 81], has received significantly less attention. We show that behavioral dynamics can result in a wide range of R_0s even at a fixed endemic prevalence and a fixed fraction of transmissions from early infection. These variable R_0s, in turn, cause elimination through UT&T to range from easy to realistically impossible. In particular, we explore conditions where, even though a relatively high fraction of transmissions come from early HIV infection (EHI), transmission can still be stopped with a relatively low effective treatment rate, because only a small fraction of transmissions must be interrupted in order to get the system below the endemic threshold of $R_0 = 1$. This accords with and extends existing modeling work that found that the relative transmissibility during EHI has little effect on the expected reduction in incidence thirty years after the initiation of an antiretroviral therapy (ART) intervention[36, 101].

In this chapter we seek to make the complex phenomena behind these relationships more understandable, so they can better inform policy decisions. This chapter seeks to generate qualitative understandings that will ultimately advance quantitative assessments. Toward this end, we use deterministic compartmental models (DCMs) to model population dynamics. Such models do not account for stochastic features of real-world population dynamics, but they allow for a simpler analysis.

Although a DCM does not directly model individuals, it can be understood as the limiting case of an individual-based model of an arbitrarily large population, and flows between compartments can therefore be conceptualized in terms of changes in the states of (large numbers of) individuals. We will make frequent use of this abstraction throughout this chapter (and in this dissertation generally).

We make our points by examining and explaining the behavior of a model for HIV transmission among men who have sex with men (MSM) closely based on the model developed by Zhang et al.[132] This model greatly simplifies the treatment cascade and the natural history of infection, while at the same time adding details about temporal patterns of behavior that other models often omit. In order to focus on the effects of these dynamics, we simplify the natural history of infection to a relatively brief early phase (early HIV infection, EHI) followed by a longer chronic phase. In Section A.5.4, we show that our qualitative results continue to hold if this simplified natural history is replaced with a more realistic one, based on results from a study of HIV transmission in Lilongwe, Malawi[100].

Similarly, we collapse the treatment cascade into a single parameter measuring the rate at which infected individuals are tested, treated, and rendered permanently non-transmitting, which we refer to as the effective treatment rate. Each individual is therefore classified as effectively treated or untreated. The untreated category includes undiagnosed; diagnosed, but untreated; and treated, but not virally suppressed individuals. In addition, all sexual partnerships are treated as instantaneous symmetrical contacts. Although all of these simplifications will affect some of the quantitative results that we obtain, we do not believe that any of them affect our qualitative conclusions. For more details about the model used, see Section A.4.1.

The purpose of these simplifications is to highlight the effect that different pat-

terns of contact rates over time can have on the response of transmission systems to universal test-and-treat interventions, even when other aspects of those systems are identical. Our take-home message is that populations can differ with regard to patterns of contact rates in ways that drastically affect the effort required to achieve elimination of ongoing transmission through UT&T. We illustrate this with a simplified model of HIV transmission among MSM that is designed to highlight certain aspects of behavioral complexity. From this model we gain insights that suggest other aspects of real-world transmission systems that could generate similar effects.

3.3 Results

3.3.1 Characteristics of interest

The primary characteristics of sexual contact patterns that we examine in this chapter are *risk heterogeneity* and *episodic risk*. Risk in our model is only determined by contact rates, which are in turn determined by which of two risk groups an individual belongs to. Our key results still hold with more than two risk groups, but for the sake of simplicity, we focus on a model with only two, which we will refer to as the higher risk group and the lower risk group.

In models where individuals always remain in the same risk group, we say there is static risk heterogeneity. In models where individuals transition between risk groups, we say there is episodic risk. We model episodic risk by selecting risk groups for individuals newly entering the sexually active population, and then reselecting their risk groups at intervals corresponding to a fixed rate of reselection (ω). The probability of being selected to the higher or the lower risk group is always the same. Consequently, the fraction of the population that is in the higher risk group at the disease-free equilibrium is unaffected by whether there is static risk heterogeneity or episodic risk. The average length of an individual's stay in a risk group is therefore

determined by both the re-selection rate and the probability of being assigned to that risk group at each point when they are re-randomized to risk groups. Thus, for example, if the probability of being assigned to the high risk group were 0.2, the average stay in the lower risk group would be four times the average stay in the higher risk group if no one left the sexually active population.

Note that this parameterization of the episodic risk formulation is the major way that this episodic risk model differs from the model in Chapter II[132]. Instead of rates of transitioning between risk groups, we have this single re-selection rate parameter. This parameterization is chosen because it simplifies the mathematical analysis of the system and makes the origin of episodic risk effects clearer, and because it generalizes naturally to a model with more than two risk groups. It is not a causal model of what generates episodic risk, but an abstraction that allows us to make important qualitative observations.

We parameterize differential transmissibility over the course of infection with a parameter (ζ) that gives the ratio of per-act transmissibility during early HIV infection (EHI) to per-act transmissibility during later/chronic infection. We will refer to this ratio as the relative transmissibility during EHI.

We focus most of our attention in this chapter on these two parameters (ω and ζ) in order to highlight the interaction of biological and behavioral factors in determining the difficulty of achieving elimination through UT&T, with a particular focus on an aspect of sexual behavior (episodic risk) that has received relatively little attention.

3.3.2 Homogeneous transmission potentials and episodic risk

To develop theory in a progressive fashion, we begin by considering a behaviorally homogeneous population, and we examine how the basic reproduction number changes when we add first static risk heterogeneity, and then episodic risk. In

order to achieve maximum clarity, we first do this with a single-stage SI model, and then show how the results extend naturally to models with multiple stages of disease progression (such as the one used in this chapter).

When the population is behaviorally homogeneous, and there is only one stage of infection, the basic reproduction number is simply the product of the (average) contact rate (χ), the per-contact transmissibility (β), and the average duration of infection (D). We will refer to this basic reproduction number, calculated under conditions of behavioral homogeneity, as the (total) homogeneous transmission potential, and denote it by H:

$$R_0 = H = \chi\beta D \tag{3.1}$$

This behavioral homogeneity will be achieved if the "higher" and "lower" risk groups actually have the same contact rate (χ) and there is only one transmission probability (β) over the entire duration of infection (D).

We elaborate this model by introducing static risk heterogeneity, i.e. by allowing the two risk groups to have different contact rates, but requiring that the re-selection rate be 0, so that individuals stay in the same risk group for their entire sexually active lives. We assume here that risk groups mix proportionately[88], except where otherwise noted. In Section A.4.6, we analyze a model of assortative mixing between risk groups and show that our major qualitative results still hold. Given static risk heterogeneity, the basic reproduction number increases in proportion to the square of the coefficient of variation of contact rates (equation adapted from May and Anderson[85]:

$$R_0 = H(1 + c_V(\chi)^2) \tag{3.2}$$

This occurs because members of the higher risk group will have both an elevated propensity to become infected and (once infected) an elevated propensity to transmit. Note that Equation 3.2 depends on the average contact rate (which determines H) and on the coefficient of variation of contact rates, but not on other features of the distribution of contact rates. Consequently, it still holds if the model is extended to have more than two risk groups, or to have a continuous distribution of contact rates.

We now further elaborate the model by introducing episodic risk, i.e. by allowing the re-selection rate to be greater than 0. In this formulation, an individual who is in the higher risk group at the time of infection will now have an elevated propensity to transmit only until their risk group is re-selected. Because the (multiplicative) increase to the basic reproduction number is a result of the same individuals who have an elevated propensity to become infected also having an elevated propensity to transmit, it will therefore only be applied to that fraction of the homogeneous transmission potential which represents transmissions occurring before we re-select the contact rate of the transmitter. This is simply the fraction of time infected that is, on average, spent prior to our re-selection of the infected individual's contact rate. We denote that fraction (which we call the (overall) fraction of heterogeneity effect) by ψ. This leads to the following generalization of Equation 3.2:

$$\begin{aligned} R_0 &= (1-\psi)H(1) + \psi H(1 + c_V(\chi)^2) \\ &= H(1 + \psi c_V(\chi)^2) \end{aligned} \tag{3.3}$$

Now we go back to the situation without risk heterogeneity and examine the situation with multiple stages of infection, each with its own per-contact transmission rate. In this case the total homogeneous transmission potential (and therefore

the basic reproduction number) will be the sum of the homogeneous transmission potentials from each stage (H_i):

$$R_0 = H = \sum_i H_i = \sum_i \chi \beta_i D_i \tag{3.4}$$

Note that in this case D_i is the average duration of stage i for all infected individuals, including (for later stages) those who do not survive long enough to enter stage i (i.e., it incorporates the probability of not reaching stage i at all). Although the focus of this chapter is on a model with only two stages of infection (and therefore the above summation is from $i = 1$ to $i = 2$), Equation 3.4 does not depend on that fact, nor do any of the subsequent equations in this section. Consequently, they will still hold if a model with a more realistic natural history of infection is used, as is the case in Section A.5.4.

Given static risk heterogeneity, the contributions of each stage of infection to the basic reproduction number will be increased by the same multiplicative factor (relative to behavioral homogeneity), and so Equation 3.2 still applies. But when we introduce episodic risk, this will no longer be the case: Because the stages of infection come in order, one after the other, the average fraction of an infected individual's time spent in the i-th stage of infection that occurs before his contact rate is re-selected (the fraction of heterogeneity effect for stage i, which we will denote ψ_i) will be smaller for larger i. In particular, the fraction of heterogeneity effect for chronic infection (ψ_2) will always be less than the fraction of heterogeneity effect for EHI (ψ_1). Therefore, the basic reproduction number will now be:

$$R_0 = \sum_i H_i(1 + \psi_i c_V(\chi)^2) \tag{3.5}$$

Equation 3.3 still applies; however, the overall fraction of heterogeneity effect (ψ) is no longer simply the fraction of an individual's time infected that occurs prior

to the re-selection of his contact rate, because one month of chronic infection and one month of EHI do not contribute equally to the total homogeneous transmission potential (H). Consequently, the overall fraction of heterogeneity effect is now a weighted average of the fractions of heterogeneity effect for each stage:

$$\psi = \frac{\sum_i \psi_i H_i}{\sum_i H_i} \tag{3.6}$$

Combining Equation 3.6 with Equation 3.3, we obtain two major implications: First, the more episodic that risk heterogeneity is (i.e. the higher the re-selection rate), the smaller ψ will be, and therefore the lower R_0 will be. Second, because the fraction of heterogeneity effect during EHI (ψ_1) will always be greater than the fraction of heterogeneity effect during chronic infection (ψ_2), episodic risk reduces transmissions from EHI less than it reduces transmissions from chronic infection. Consequently, episodic risk raises the fraction of transmissions from early infection while it lowers the basic reproduction number.

The above analysis relates to the potential for transmissions during exponential growth. Under these conditions, when the prevalence of infection is negligible in both risk groups, episodic risk has a strictly negative effect on transmissions. The effects of episodic risk on transmissions when the prevalence is not negligible (such as at the endemic equilibrium), are more complex, as seen in Section 3.3.3.

3.3.3 Endemic prevalence and the basic reproduction number as functions of the relative transmissibility during EHI and the re-selection rate

In Figure 3.1, we illustrate the effects of the relative transmissibility during EHI (ζ) and the re-selection rate (ω) on the basic reproduction number (R_0) and the endemic prevalence (P), while fixing the total transmission potential across the course of infection, and keeping most other parameters the same (full details in Section A.4.4.

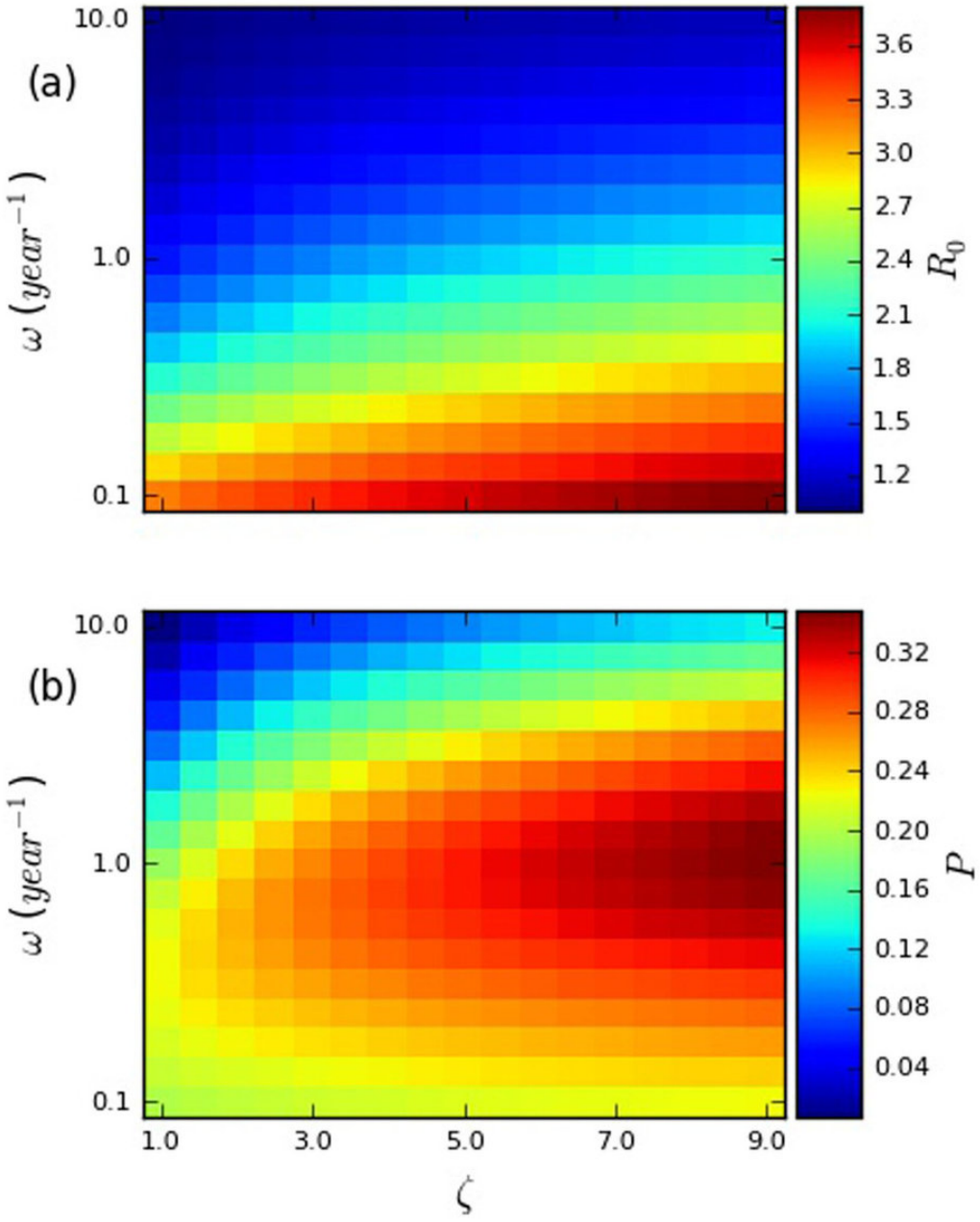


Figure 3.1: Heatmaps plotting (a) the basic reproduction number (R_0), and (b) equilibrium prevalence (P) against the relative transmissibility during early HIV infection (ζ) and re-selection rate (ω). The per-act transmissibilities during EHI and chronic infection for each parameter set were chosen based on the constraints that (1) their ratio must be ζ and (2) the total transmission potential in a homogeneous system must be the same for all parameter sets. This transmission potential was chosen by setting the endemic prevalence for the lower-left parameter set to be 0.2. The other parameters used are summarized in Table A.1. Each panel has a scale that runs from the lowest to the highest value observed in that panel.

In Figure 3.1, both the basic reproduction number (R_0) and the endemic prevalence (P) are monotonically increasing with respect to the relative transmissibility during EHI, even though the overall transmission potential is fixed. This is a consequence of the fact, noted above, that episodic risk reduces the boost in transmissions caused by risk heterogeneity less for transmissions during EHI than for transmissions during chronic infection. Consequently, if the total (homogeneous) transmission potential is fixed, increasing transmissions from EHI at the expense of transmissions from chronic infection results in an increase in transmissions, leading to an increase in both R_0 and P.

Unlike the relative transmissibility during EHI, the re-selection rate has a drastically different effect on the basic reproduction number and the endemic prevalence. As noted in Section 3.3.2, the basic reproduction number is monotonically decreasing with respect to the re-selection rate, because re-selection attenuates the increase in the average contact rate of infected individuals that results from risk heterogeneity. In contrast, the endemic prevalence has a non-monotonic dependence on the re-selection rate, peaking at a relatively high value of the latter. This is a product of the combined effects of the attenuation of the average contact rate of infected individuals (which works to reduce the endemic prevalence, just as it does the basic reproduction number) and the increased replenishment of higher risk susceptibles, which works to increase the endemic prevalence[132].

That episodic risk can increase endemic prevalence relative to the basic reproduction number at a given observed degree of (instantaneous) risk heterogeneity may explain the difficulty that modelers have experienced in reconciling estimates of the coefficient of variation in contact rate and the basic reproduction number with observed prevalence of HIV infection among MSM[7].

3.3.4 System determinants of key epidemiological measures

We now introduce a third outcome of interest: the effective treatment rate required to achieve elimination (τ_E). As noted in the introduction, we collapse the treatment cascade into a single parameter, which we refer to as the effective treatment rate and denote τ. We then define the effective treatment rate required to achieve elimination (τ_E) as the minimum value of the effective treatment rate (τ) at which elimination is (deterministically) achieved, i.e. the minimum value that reduces the basic reproduction number (R_0) to or below the endemic threshold of $R_0 = 1$. In practice, achieving elimination in a reasonable timeframe may require that R_0 be brought significantly below 1. Consequently, if this model or one like it were to be used to set goals for a particular UT&T campaign, it would likely be desirable to target a lower basic reproduction number. Rather than fixing realistic targets, here we merely seek to provide qualitative insights.

In Figure 3.2, we consider the effects on the basic reproduction number (R_0), fraction of transmissions from early infection (ϕ), and effective treatment rate required to achieve elimination (τ_E) of the same parameters examined in Figure 3.1, the relative transmissibility during early infection (ζ) and the re-selection rate (ω). In this case, however, instead of fixing the total transmission potential over the course of infection, we fix the endemic prevalence, at 0.2. We vary the total transmission potential (by varying β_1 and β_2, proportionally) in order to achieve this fixed endemic prevalence. This provides the common perspective in risk factor and prevention effects for HIV infection, where the prevalence is known, but what is determining that prevalence is unknown.

An important aspect of these relationships is that both factors have strong effects on the effective treatment rate required to achieve elimination as shown in panel (a)

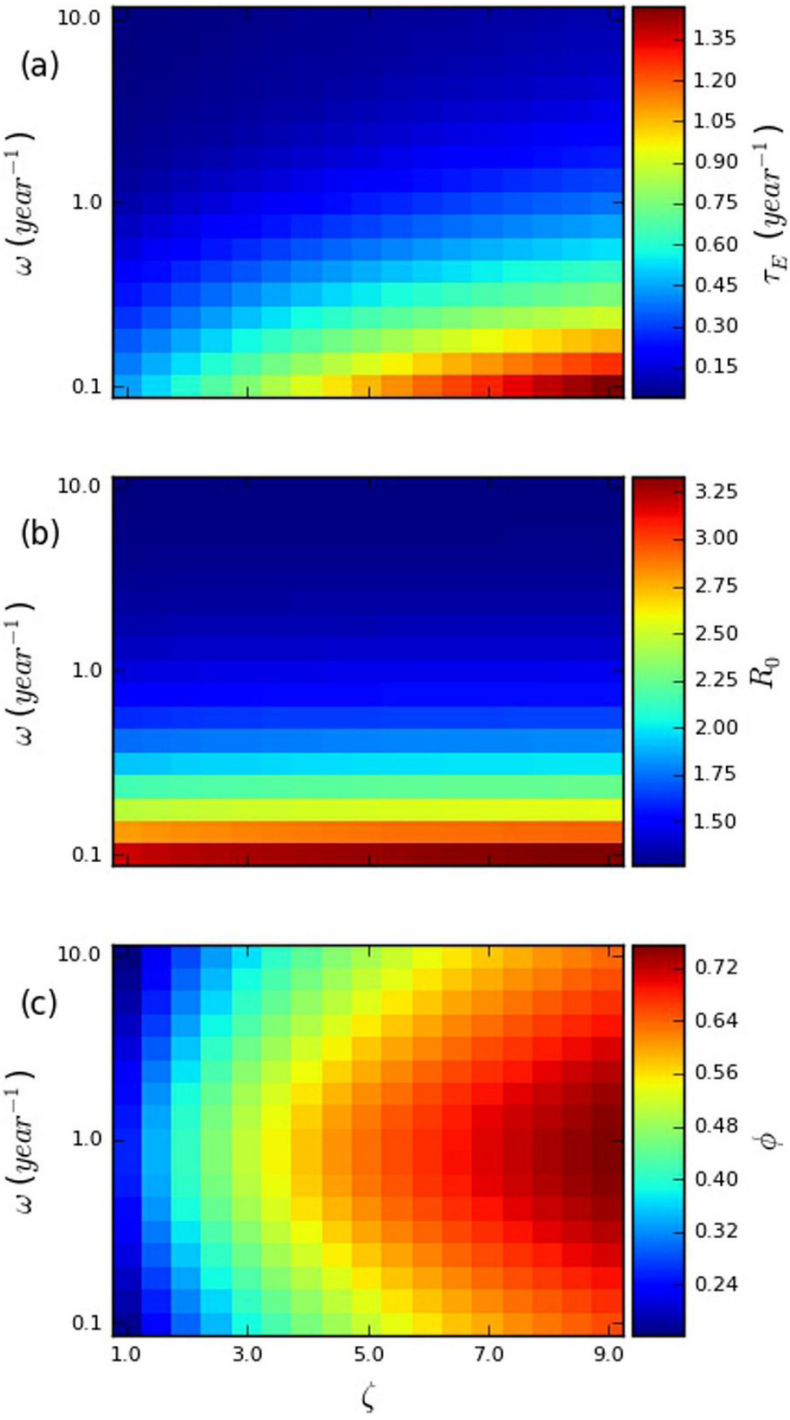


Figure 3.2: Heatmaps plotting (a) the effective treatment rate required to achieve elimination (τ_E), (b) the basic reproduction number (R_0), and (c) the fraction of transmissions from early HIV infection (ϕ) against the relative transmissibility during early HIV infection (ζ) and re-selection rate (ω). The other parameters used are summarized in Table A.1. Each panel has a scale that runs from the lowest to the highest value observed in that panel.

of Figure 3.2. Increasing the re-selection rate from 0.1 per year (relatively static risk groups) to 10.0 per year (strongly episodic risk) while holding the relative transmissibility during early infection constant at 9 results in a reduction of the required effective treatment rate from 1.47 to 0.101. This is over a 14-fold difference, without any change in the biological parameters, nor in any behavioral parameters that could be readily measured cross-sectionally. Likewise, reducing the relative transmissibility during early infection from 9 to 1 (while holding the re-selection rate constant at 0.1) results in a reduction of the necessary effective treatment rate to 0.442 – over a three-fold difference.

To understand what is generating these effects, it is useful to understand how the effective treatment rate required to achieve elimination (τ_E) is affected by changes in the basic reproduction number (R_0, shown as the outcome in panel (b) of Figure 3.2) and the fraction of transmissions from early infection (ϕ, shown as the outcome in panel (c) of Figure 3.2). As can be seen in Figure 3.2, the effect on τ_E of the parameters ω (the re-selection rate) and ζ (the relative transmissibility during EHI) is essentially a combination of their effects on R_0 and ϕ. This combination of effects is attributable to the fact that, as mentioned in the introduction, R_0 and ϕ each represent a different aspect of difficulty in eliminating transmissions through UT&T: The higher R_0 is, the larger the fraction of transmissions that must be prevented, and the higher ϕ is, the less effective UT&T is at preventing transmissions. A more detailed consideration of how R_0 and ϕ affect τ_E can be found in Section A.5.1.

In panel (c) of Figure 3.2, we can see that the fraction of transmissions from EHI (ϕ) is, unsurprisingly, strongly positively dependent on the relative transmissibility during EHI (ζ). What is more striking is the strong, but non-monotonic, dependence on the rate of risk-group re-selection (ω), peaking when $\omega < 1$/year. The explanation

for this phenomenon is that the fraction of transmissions from EHI is maximized when the difference in average contact rates between EHI and chronic infection is maximized, and this happens at a moderate re-selection rate: At very low re-selection rates, there is little difference, because the vast majority of infected individuals still have the same contact rate as when they were infected, regardless of their stage of infection. At very high re-selection rates, there is again little difference, because the vast majority of infected individuals have had their contact rate re-selected since they were infected, again, regardless of their stage of infection. It is only at intermediate re-selection rates that individuals in chronic infection are substantially more likely to have had their contact rates re-selected since being infected than individuals in EHI. Further details are included in Section A.5.2.

3.3.5 Endemic prevalence as a function of R_0

In order to illustrate the importance of various aspects of the model in determining the basic reproduction number R_0, we plotted curves relating R_0 to the endemic prevalence (P) for several related models. We derived a maximal model (parameter set given in Table A.1) from our primary model by adding assortative mixing ($m = 0.5$), in which individuals in a particular risk group are more likely to form contacts with other individuals in that risk group[66, 88], to our primary model. We derived static risk heterogeneity sub-models from both the maximal model and our primary model by setting the re-selection rate (ω) to 0. We further derived a behavioral homogeneity model by setting the ratio of the contact rates of the two risk groups (r_{HL}) to 1. Further details are included in Section A.4.5. The results are shown in Figure 3.3.

It is particularly instructive to consider what effect transitioning between these submodels has when the prevalence is fixed, as it is in Figures 3.2 and 3.4. Because

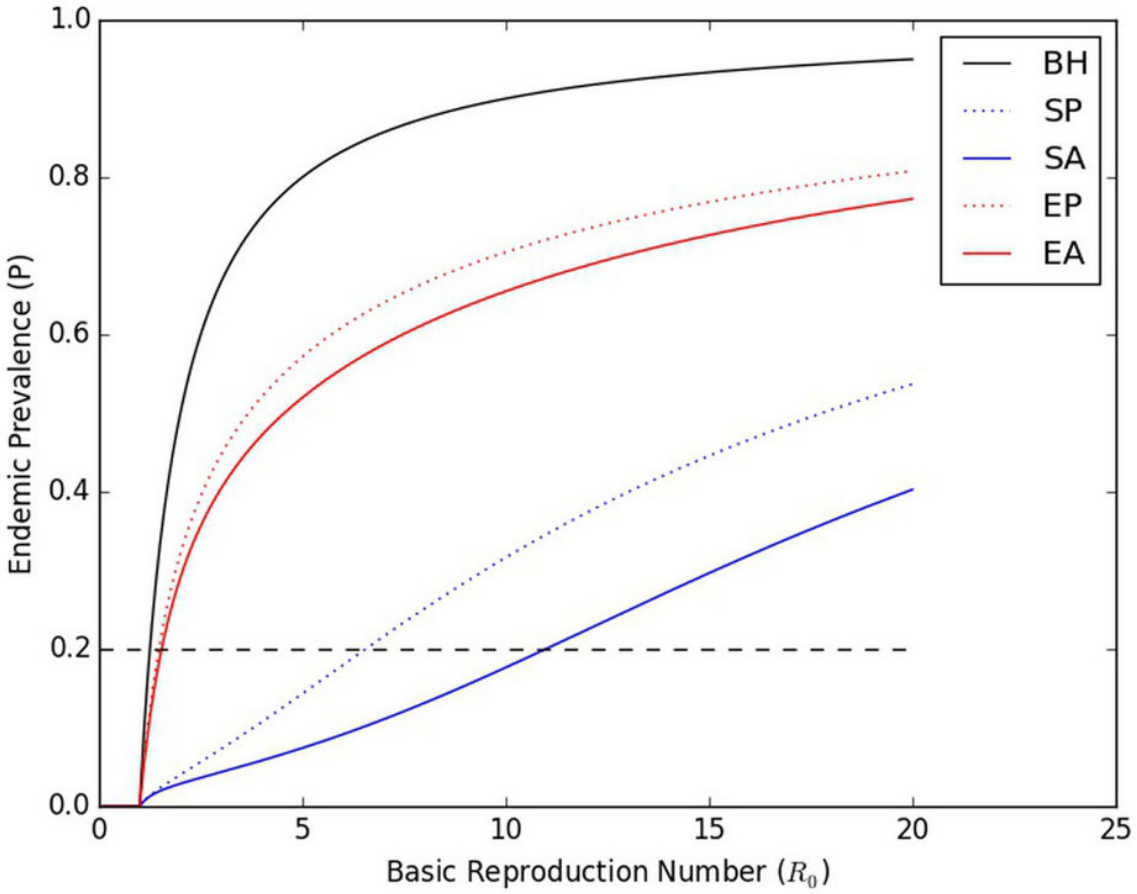


Figure 3.3: Curves showing the prevalence as a function of R_0 when all parameters except the overall transmissibility are held constant. The curves shown are for the full (maximal) model (episodic risk, with assortative mixing ($m = 0.5$) (EA)), and several reduced models: Episodic risk, with proportional mixing (EP, the primary model in this chapter); Static risk heterogeneity, with assortative mixing ($m = 0.5$) (SA); Static risk heterogeneity, with proportional mixing (SP), and Behavioral Homogeneity (BH). The formulation of these reduced models is discussed in more detail in Section A.4.5. The dashed black line indicates a constant prevalence of 0.2, illustrating the drastically different values of R_0 that are possible when the prevalence is fixed.

Symbol	Meaning
L	Total population of lower risk individuals
S_{L}	Susceptible, lower risk
$A_{\mathrm{L,U}}$	Early-infected, lower risk, untreated
$A_{\mathrm{L,T}}$	Early-infected, lower risk, treated
$C_{\mathrm{L,U}}$	Chronically infected, lower risk, untreated
$C_{\mathrm{L,T}}$	Chronically infected, lower risk, treated
H	Total population of higher risk individuals
S_{H}	Susceptible, higher risk
$A_{\mathrm{H,U}}$	Early-infected, higher risk, untreated
$A_{\mathrm{H,T}}$	Early-infected, higher risk, treated
$C_{\mathrm{H,U}}$	Chronically infected, higher risk, untreated
$C_{\mathrm{H,T}}$	Chronically infected, higher risk, treated

Table 3.1: Symbols representing (sub)populations

Figure 3.3 depicts prevalence as a function of R_0, this amounts to selecting a horizontal line (such as the black dashed line, indicating a prevalence of 0.2), and examining how transitioning between curves changes the value of R_0 at which the curve intersects that horizontal line. By doing so, we can see how drastically an estimated R_0 can vary at a single prevalence, depending on the model used to estimate it. Failing to account for risk heterogeneity at all, or for assortative mixing, can result in a drastic underestimate of R_0, while failing to account for episodic risk can result in a drastic overestimate. Such an error can easily result in incorrect inferences about the feasibility of achieving control or elimination with a given intervention.

Consistent with the discussion in Section 3.3.3, static risk heterogeneity dramatically increases R_0 at a given prevalence, but this increase is reduced (though not eliminated) when episodic risk is introduced. Assortative mixing increases the basic reproduction number at a given prevalence for both static risk heterogeneity and episodic risk, by causing the most frequently transmitting infected individuals (those in the higher risk group) to transmit preferentially to higher risk susceptibles.

It has previously been observed that static risk heterogeneity and assortative mixing both decrease prevalence at a given R_0[6, 53]. To this we now add that episodic

Symbol	Unit	Value(s)	Meaning
χ	1/year	20	Average contact rate
γ_1	1/year	1	Rate of transition from EHI to chronic infection
γ_2	1/year	1/8.5	Rate of AIDS-related death or departure from the sexually active population during chronic infection, if untreated. Chosen to give a mean total duration of infection of 9.45 years[13].
γ_{2T}	1/year	0	Rate of AIDS-related death or departure from the sexually active population during chronic infection, if treated. Does not actually affect any results presented in this chapter.
μ	1/year	1/40.28	Rate of death or departure from the sexually active population unrelated to HIV infection; because the population size is normed to be 1 at the disease-free equilibrium, this is also the absolute rate of entry into the sexually active population. From Supplementary Text S3 of a paper by Volz *et al.*[128].
τ	1/year	Variable	Effective treatment rate
r_{HL}	-	Variable	Contact rate ratio between higher risk and lower risk individuals
f_H	-	Variable	Fraction of the population that is at higher risk at the disease-free equilibrium.
f_L	-	$1 - f_H$	Fraction of the population that is at lower risk at the disease-free equilibrium.
χ_L	1/year	$\frac{\chi}{1+(r_{HL}-1)f_H}$	Contact rate for lower risk individuals
χ_H	1/year	$r_{HL}\chi_L$	Contact rate for higher risk individuals
m	-	Variable (0–1)	Fraction of an individual's contacts that are reserved for members of the same risk group. Throughout this chapter, it is 0 unless stated otherwise.
ω	1/year	Variable (0.01–10.00)	Rate at which we randomly re-select which risk group an individual is in (re-selection rate)
ω_H	1/year	$f_L\omega$	Rate at which higher risk individuals transition to the lower risk group
ω_L	1/year	$f_H\omega$	Rate at which lower risk individuals transition to the higher risk group

Table 3.2: Symbols representing parameters (including derived parameters), part 1: Parameters not involving or dependent on transmissibilities

Symbol	Unit	Value(s)	Meaning
ζ	-	Variable (1.0–9.0)	Relative transmissibility during EHI
β_1	transmissions/ contact	$\zeta\beta_2$	Per-contact transmissibility during EHI
β_2	transmissions/ contact	[used to fit the model to a target prevalence or basic reproduction number]	Per-contact transmissibility during chronic infection
H_1	transmissions	$\frac{\chi\beta_1}{\mu+\gamma_1+\tau}$	The average number of secondary transmissions during EHI per index case, under behavioral homogeneity.
H_2	transmissions	$\frac{\chi\gamma_1\beta_2}{(\mu+\gamma_1+\tau)(\mu+\gamma_2+\tau)}$	The average number of secondary transmissions during chronic infection per index case, under behavioral homogeneity.
H	transmissions	H_1+H_2	The average number of secondary transmissions, under behavioral homogeneity.
ψ_1	-	$\frac{\mu+\gamma_1+\tau}{\mu+\gamma_1+\tau+\omega}$	The average fraction of time spent in untreated EHI that occurs prior to the first re-selection of that individual's contact rate after his infection.
ψ_2	-	$\frac{(\mu+\gamma_1+\tau)(\mu+\gamma_2+\tau)}{(\mu+\gamma_1+\tau+\omega)(\mu+\gamma_2+\tau+\omega)}$	The average fraction of time spent in untreated chronic infection that occurs prior to the first re-selection of that individual's contact rate after his infection.
ψ	-	$\frac{\psi_1 H_1+\psi_2 H_2}{H_1+H_2}$	The average fraction of secondary transmissions (under behavioral homogeneity) that occur prior to the first re-selection of that individual's contact rate after his infection.

Table 3.3: Symbols representing parameters (including derived parameters), part 2: Parameters involving or dependent on transmissibilities

risk increases prevalence at given R_0, relative to static risk heterogeneity. In fact, all three of these observations reflect different manifestations of the same basic phenomenon: Provided that the dynamics of behavior and disease are independent (e.g. there is neither serosorting nor faster disease progression among certain behavioral groups), static risk heterogeneity and assortative mixing will decrease prevalence at a given R_0 to the extent that they cause transmissions to disproportionately occur into a subgroup that saturates with infection more rapidly than the population as a whole.

They do this by concentrating transmissions into the higher risk group, which consequently can become saturated with infection even when the prevalence in the population as a whole is still quite low. Episodic risk (partially) counteracts this effect (and thus increases prevalence at a given R_0) by causing a net replacement of higher risk infecteds by (formerly) lower risk susceptibles, thereby reducing the saturation of the higher risk group, and by moving infected individuals into the lower risk group, where assortative mixing will increase the fraction of their contacts that are with susceptible individuals.

3.3.6 Re-selection rate effects when the fraction of transmissions from EHI (ϕ) is fixed

In Figure 3.4, we illustrate the dramatic effects that the re-selection rate can have on the ease of elimination, even when the fraction of transmissions from early infection is fixed. In order to obtain the results in Figure 3.4, we fixed the endemic prevalence at 0.2 and the fraction of transmissions from early infection (ϕ) at 0.447, varying the per-contact transmission rate during chronic infection (β_2) and the relative transmissibility during EHI (ζ) in order to do so. This value for ϕ was chosen based on the results of a recent phylodynamic analysis[128]. In this figure, changing

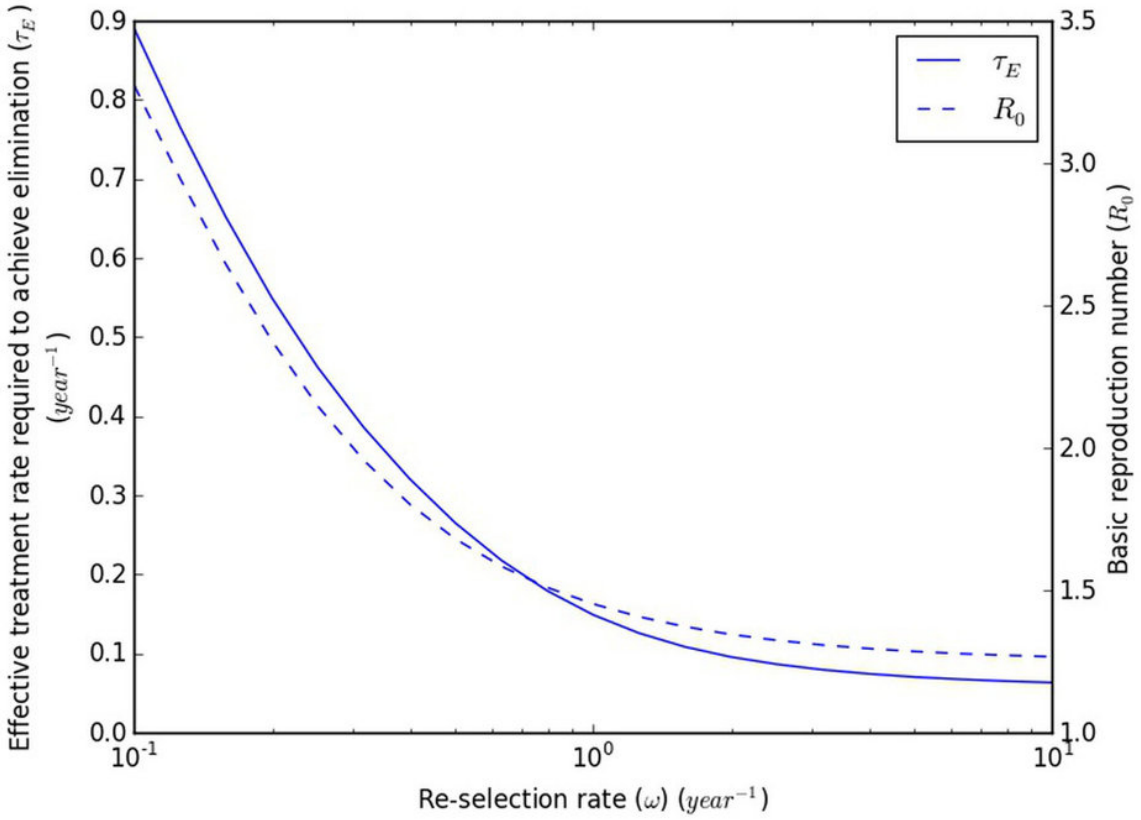


Figure 3.4: The basic reproduction number (R_0) and effective treatment rate required to achieve elimination (τ_E) as a function of the re-selection rate (ω). The prevalence is fixed at 0.2, and the fraction of transmissions from EHI (ϕ) is fixed at 0.447[128]. To achieve this, the transmissibilities from each of acute and chronic infection are allowed to vary; all other parameters are fixed. Details are presented in Section A.4.7.

the re-selection rate can be seen to result in R_0 values that range from 1.27 to 3.28 and τ_Es that range from 0.0635/year to 0.891/year.

To get a sense of the magnitude of the latter difference, let us suppose (1) that time from infection to diagnosis, time from diagnosis to linkage to care, and time from linkage to care to viral suppression are all exponentially distributed, and (2) that the effective treatment rate is the rate of (also exponentially distributed) single-step diagnosis, treatment, and viral suppression that generates a distribution of times from infection to suppression that has the same mean as the actual distribution. Using published estimates for the time from diagnosis to linkage to care (77.2% initiating

care with 3 months)[124] and the time from linkage to care to suppression (median of 1.03 years for all patients)[23], we find that a τ_E of 0.0635/year corresponds to a diagnosis rate of only 0.0710/year. In contrast, a τ_E of 0.891/year would not be achievable at all with current times from linkage to care to viral suppression (median of 0.87 years even when restricted to patients with perfect retention in care)[23], even if both diagnosis and linkage to care were instantaneous, and retention in care were perfect. Details of this calculation are presented in Section A.5.3.

From these results, we can see that inferences about the feasibility of achieving control or elimination through UT&T are extremely dependent on assumptions about if and how individuals' contact rates change over the course of their lives, even when the prevalence, fraction of transmissions from EHI, and (instantaneous) distribution of contact rates are all known.

3.4 Discussion

We have shown that episodic risk can have a dramatic effect on the effort required to achieve elimination through treatment, even when both the prevalence and the fraction of transmissions from early infection are fixed. Moreover, we have shown that episodic risk can have a major effect on the fraction of transmissions from early infection itself, independent of biological factors that increase transmission during EHI. We have also shown how differences in the extent to which risk heterogeneity is episodic can make the difference between a transmission system that is extremely vulnerable to universal test and treat, and one for which no level of testing and treating can ever be sufficient, by itself, to eliminate ongoing HIV transmission. Others have illustrated how transmission during early infection can affect the relationship between prevalence and the basic reproduction number, and thus alter the popula-

tion effects of early diagnosis and treatment, sometimes in unexpected ways[36, 101]. We have illustrated here how episodic risk is likely to considerably amplify those effects. Consequently, collecting more data on risk heterogeneity and episodic risk (or its continuous generalization, risk volatility[110]) is highly important in order to make the best possible decisions about how to allocate finite HIV control resources.

Our models do not attempt to reproduce real-world transmission dynamics. They dichotomize what are actually continuous phenomena: transmissibility over the course of untreated infection, contact rates, change in contact rates, and treatment success or failure. They are kept simple in order to focus on the basic concepts of how changes in contact rates over the course of individuals' lives alter the fraction of transmissions from early infection, R_0, and the effectiveness of UT&T. This understanding should affect how control decisions are pursued in the face of ignorance about how individuals change their contact rates over the course of their lives. It should also affect both field studies and modeling studies designed to reduce our ignorance. The simplicity of our models, however, means that any attempt to predict the effectiveness of UT&T in a particular population based on our results is invalid in the absence of an assessment of how realistic relaxation of our simplifying assumptions affects our results. However, we have identified important qualitative phenomena which will be highly relevant even in more realistic models.

In this chapter, we have presented episodic risk abstractly, without considering its determinants. As noted in Chapter II[132], there are numerous causal phenomena that can result in periods of more frequent or riskier sexual activity than characterizes an individual's sexually active life as a whole, such as periods of recreational drug use or alternation between a partnered state (with a low or zero rate of outside sexual contacts) and an unpartnered state (with a higher rate of casual sex). Higher rates

of sexual activity at certain ages can also produce similar phenomena, particularly when sexual activity shows assortative mixing by age, as is generally the case. Some of these effects have been captured in previous model analyses that have explored the role of EHI on transmission in populations[65, 78]. But empirical work on these factors has not progressed. We hope that our simple way of abstracting risk behavior fluctuations with our episodic risk formulation helps to show why risk fluctuations are so important. But we are not advocating that the episodic risk abstraction should be the way that histories are obtained. How risk fluctuation histories are taken should depend on what questions are easiest to comprehend and validly answer and how well the answers to those questions allow researchers to measure changes in risk behavior over the course of an individual's life.

One consistent pattern that we have observed across various parameters that we have examined is that changes to parameters tend to make elimination through UT&T harder when they create or strengthen the phenomenon of a distinct core group that transmits heavily to itself. That this phenomenon is consistent across various ways of strengthening such a group (increasing contact rate heterogeneity, decreasing episodic risk, and increasing assortativity) suggests that it is likely to be robust to realistic relaxation of simplifying assumptions. Under these conditions, UT&T becomes less effective. These conditions, however, form a core group with persistent higher risk behavior where Pre-Exposure Prophylaxis is more practicable and effective.

One possible direction for further research is extending the various dichotomized aspects of the model to polychotomous or continuous measures. For example, dichotomized episodic risk can be replaced by continuous risk volatility[109, 110]. Another is incorporating important aspects of sexual behavior that are not included

in the present model, such as long-term monogamous or semi-monogamous partnerships, temporal changes in insertive or receptive behavior[5] or temporal changes in condom use with different types of partners.

Episodic risk behavior/risk volatility has been understudied. Two studies have found both between and within-individual volatility[107, 108]. However, due to limitations in the datasets that were available, these were based on an analysis of data from the early 1990s, and therefore may not be reflective of current behavioral dynamics.

There are three fundamental messages that we hope that other researchers will take from this chapter. First, we have presented a new methodology for calculating the basic reproduction number in the presence of episodic risk, which generalizes naturally to continuous risk volatility, and which helps in creating understanding not only of the reasons for the effects of episodic risk themselves, but also of their interactions with other behavioral or biological phenomena. We hope that this approach will be of service to other modelers. Second, we have shown how episodic risk can increase prevalence at the same time as it reduces the basic reproduction number and the difficulty of eradication. Third, we have shown how strong these effects can be. Consequently, they should be taken into account, and the necessary data gathered in order to determine their magnitude and better plan control activities using models that incorporate the strong effects we have illustrated.

3.5 Methods

Throughout this chapter, all sexual behavior is modeled as instantaneous, symmetric contacts. All sexual contacts between an infected individual and a susceptible individual are assumed to have a potential for HIV transmission that depends only

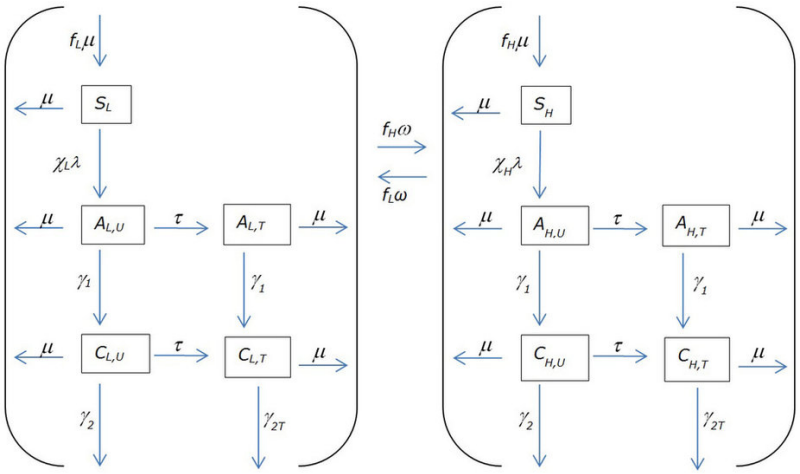


Figure 3.5: Dynamics of acquisition, progression, and treatment of infection. The meanings of symbols representing subpopulations are given in Table 3.1, and the meanings of symbols representing parameters (including derived parameters) are given in Tables 3.2 and 3.3. Note that only re-selections resulting in transition to the other risk group are shown; loops depicting re-selections that place the individual in the same risk group (and therefore same compartment) they were in prior to re-selection are not included.

on the stage and treatment status of the infected individual. Per-contact transmissibility is in general higher during EHI than during chronic infection, with the ratio between the two transmissibilities being a parameter whose effects we explore.

The flow diagram for our primary model is depicted in Figure 3.5. More details, parameter definitions and ranges, and a system of coupled differential equations are presented in Section A.4.1.

All simulations were performed and numerical results obtained using the Anaconda distribution (Continuum Analytics, version 2.1.0, 64-bit) of Python (version 3.4.1). Apart from the standard library, we made use of the NumPy (version 1.9.0) and SciPy (version 0.14.0) packages for simulation and computation, and the matplotlib (1.4.0) package for visualization and graphics production. All integration of systems of ordinary differential equations (ODEs) was done using the scipy.integrate.odeint function, with default solver and stepsize. Further details may be found in the source code. Means of obtaining that code are discussed in Appendix

B.

In general, we have three outcomes of interest: the basic reproduction number (R_0), the fraction of transmissions from EHI (ϕ), and the effective treatment rate that is necessary to completely eliminate ongoing transmission (τ_E). For the purposes of this chapter, we use the next-generation matrix (NGM) definition of R_0 developed by Diekmann et al.[25] We define ϕ as the fraction of transmissions from EHI made by all individuals in a generation during exponential growth (under the same assumptions used to define the next-generation matrix), and solve for it using a modification of the methods for finding R_0[25], discussed in Section A.4.6. Using these definitions, we obtain an algebraic equation giving τ_E in terms of R_0 and ϕ for sub-models without episodic risk. This equation is presented, along with its implications, in Section A.4.6.

Except where otherwise noted, the endemic prevalence in the absence of treatment was set to equal 0.2, and for each parameter set, a value of β_2 that generated that endemic prevalence was obtained numerically. All of the software listed above is free to use. Directions for obtaining a copy of the source code are included in Appendix B.

CHAPTER IV

Dynamics Affecting the Risk of Silent Circulation when Oral Polio Vaccination is Stopped

4.1 Chapter preface

In this chapter and Chapter V (originally the main text and supplementary material of a paper, respectively[77]), I present and analyze a deterministic model of poliovirus transmission designed to highlight the role of ongoing waning in the potential for poliovirus transmission systems to generate sustained silent circulation. This model includes strain-specific characteristics, variable speed and depth of waning, and different histories of eradication efforts. Using this model, I show that uncertainty about waning dynamics translates to uncertainty about the potential for silent circulation lasting longer than the 3 year criterion used by the GPEI to certify elimination of circulating poliovirus from a country, creating a risk that elimination will be certified in error, with potentially disastrous consequences. I then present recommendations for future research designed to mitigate this risk.

4.2 Introduction

The final stages of polio eradication are approaching. The eradication campaign began in 1988 and has lasted almost 30 years. During that time, it encountered and overcame many foreseen and unforeseen challenges. During the first couple decades

of the eradication program, detecting paralytic cases through acute flaccid paralysis surveillance (AFPS)[51, 60] was sufficient to direct vaccination efforts to where they were most needed. Paralytic poliomyelitis cases occur almost exclusively after first infections. Transmissions between people who have been either previously infected or successfully vaccinated are silent in the sense that they are unlikely to be detected by AFPS. Such transmission events have not to date been an obstacle to eliminating polio from most of the world. However, the few countries where elimination has yet to be achieved are unique in their conditions of high transmissibility and long periods of inadequate vaccination. These conditions could alter how long silent circulation (SC) can be sustained. SC begins with the last AFPS case detected and ends either with a new AFPS case detection or complete eradication of all infections.

Using a transmission model, we develop theory that relates SC duration to waning immunity dynamics, transmissibility of both oral polio vaccine (OPV) and wild polio virus (WPV), and vaccination rates over time. We clarify how and why these factors may extend SC beyond the current three year criteria used to define when a country is considered polio-free. The new insights we generate contribute to a theoretical framework that can help guide decisions regarding what data should be used to decide when to stop OPV use and what actions will minimize SC when OPV use is stopped.

The potential for prolonged SC is currently an issue because the final polio eradication plan requires cessation of all oral polio vaccine (OPV) use[64]. The need to cease OPV use arises because the vaccine strain can cause paralysis and can be transmitted and evolve to cause paralysis and to transmit as readily as wild polio viruses (WPV) [20, 41]. OPV cessation must be coordinated worldwide because of the risks that introduction of OPV from an area using it to an area not using it will

lead to sustained OPV transmission and evolution into circulating vaccine derived polio viruses (cVDPV) that can cause epidemic polio is great[42, 64, 90]. Inactivated polio vaccine (IPV) will replace OPV. But IPV does not stimulate immunity against infection and transmission in children who have never experienced an OPV or WPV infection[92, 94]. Thus, after OPV cessation children vaccinated only with IPV will rapidly accumulate and amplify any transmission that exists[42]. If sufficient OPV is available, the setback to eradication from undetected SC after OPV cessation may be manageable even though localized reinstitution of OPV vaccination runs a high risk of generating circulating vaccine derived epidemics[42]. But the intention is to eradicate all strains of OPV because if released they could evolve into as much of a problem as WPV[91, 103]. That means destroying all OPV vaccine reserves. If SC persisted up to the point when OPV was destroyed, control potential would revert to where it was before Sabin developed OPV.

Action is therefore required to insure the chances of prolonged SC are minimal. Two different surveillance systems are being used to this end. AFPS has been highly developed during eradication and is the major surveillance system that has helped eradication efforts reduce transmission to only three countries[51, 60]. All cases of acute flaccid paralysis under age 15 are cultured for polio virus. Globally AFPS has been designed to cover every population anywhere in the world with risks of polio transmission. In crucial populations, not only AFP cases are cultured, but their contacts are cultured as well. A newer system, environmental surveillance (ES), involves collecting samples from sewage systems or other feces contaminated black waters and identifying and sequencing polioviruses [4, 10, 62]. It indicates the presence of virus in a population but does not identify source cases. It is not being designed as a comprehensive system covering all populations. A relevant question addressed in

this chapter is where ES should be established to best detect SC. Answering that question requires an understanding of what generates SC that this chapter provides. SC has two major potential causes. The cause on which the eradication effort has focused to date is inadequate AFPS. We focus in this chapter on elucidating a second potential cause of SC. It is waning immunity that increases the potential for individuals with either OPV or WPV induced immunity to become reinfected and transmit infection.

To pursue a better understanding of SC causes, we pursue a better theoretical framework for assessing the effects of waning immunity on SC. To this end we analyzed a simple model as a first step in inference robustness assessment[75]. We simplified the polio model we developed to examine success and failures of the polio eradication effort in the context of waning immunity and OPV transmissibility[86]. Model simplifications allowed us to see clearly the contributions of different groups and different parameters to the effective reproduction number. The theoretical framework this allowed us to develop has the potential to extract considerable information about waning from combined AFPS and ES data gathered during recent years. We do not have access to that data. But we hope that the understanding that we generate here will increase the capacities for risk assessment by those who do.

The major data used in modeling efforts to date are AFPS and vaccination coverage data. But we uncover an identifiability problem of AFPS data for predicting SC if there is waning immunity. We do this by illustrating how models that do and models that do not generate SC can be fitted to the same AFPS data. Immunity against paralytic polio is thought not to wane. But immunity against transmission clearly does[1, 52, 67]. A recent analysis of more direct data on waning of immunity has succeeded in developing a new theoretical framework for modeling waning

immunity[42]. Waning formulations in that paper are realistic and supported by data. But our simpler approach elucidates issues that all more realistic models must address but are hard to see in them. Our analysis seeks to uncover how and why the following interact to affect SC risks: (1) waning immunity patterns, (2) high transmission potential, and (3) delays in getting vaccination to levels needed to reduce cases to very low levels. We also seek to develop theory to predict how vaccine transmissibility and infection to paralysis ratios (IPR) affect SC. Additionally, we explore how vaccination of age groups with waned immunity before stopping OPV could help insure eradication given different waning patterns.

4.3 Methods

4.3.1 Infection transmission model

Our transmission model is illustrated by the flow diagram in Figure 4.1. It distinguishes first infections from subsequent reinfections and WPV infections from OPV infections. Waning in the model occurs in a one-step transition from recovered state R to a partially susceptible state P. Recovery from all infections leads to a single immune state R. That means immunity from WPV or OPV infections is identical, as is immunity resulting from a first infection or a subsequent infection.

4.3.2 Simplifying model assumptions

Our model analyzes only one polio serotype at a time. It uses parameters such as the fraction of first infections that result in paralytic poliomyelitis and the relative transmissibility of OPV compared to WPV, that have been used by others[31] to characterize the three serotypes of polio. It assumes homogeneous mixing and homogeneous contact rates across all compartments with paralytic cases transmitting identically to non-paralytic first infection cases. It assumes that death rates are

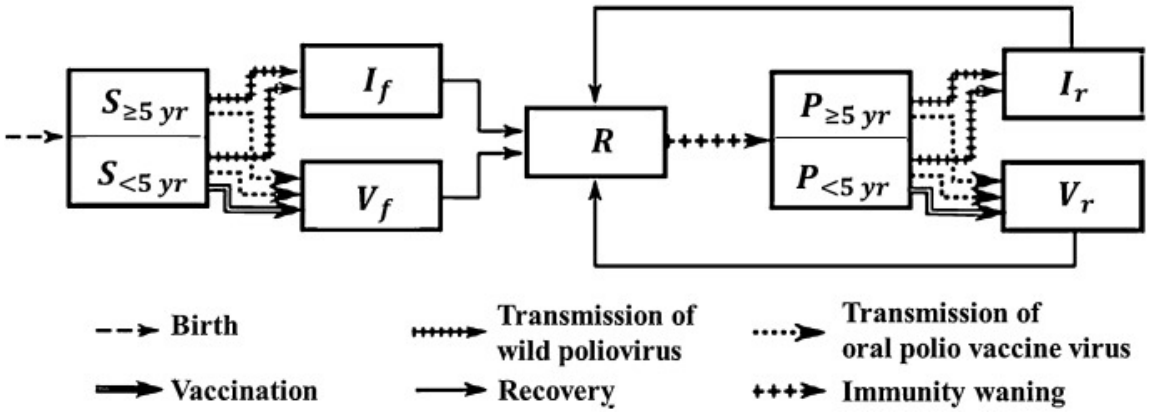


Figure 4.1: The seven compartments in our one step waning model and its flows not related to age or death are illustrated. The one waning step is from R to P. The speed of that flow is the speed of waning. The depth of waning is determined by the degree of susceptibility in the P compartment and the duration and contagiousness of the WPV or OPV infections that flow out of P. Although we distinguish between different age groups in all seven compartments (in order to allow for realistic aging), we only show here the distinctions that are relevant to infection processes, namely between (fully or partially) susceptible individuals who are or are not in the age group targeted by vaccination efforts.

constant, not affected by age, and are equal to birth rates. The differential equation model form used essentially assumes a continuous population rather than discrete individuals. It also assumes everyone in the population has an equal rate of contacting everyone else. Model equations and parameters are presented in Chapter V.

The model does not directly distinguish paralytic infections from silent infections. But our analyses of model output use the first infections to paralysis ratio (IPR). When the cumulative number of first infections in the past year becomes less than the IPR, the time of the last paralytic polio case is declared. A new count of first infections then begins and goes on without time limit. When that total count exceeds the IPR, a time of reappearance or recurrence of polio cases is declared. We assume that reinfections do not cause paralytic poliomyelitis.

Eradication is said to occur in our model when the prevalence of infection falls below 1 in a population of one million. We explored a variety of definitions that counted different entities such as numbers of first and repeat infections weighted by

their overall transmission potential. We found that none of these affected the basic conclusions of this chapter.

Different duration and contagiousness of infection are assigned to wild polio virus (WPV) and oral polio vaccine virus (OPV) infections. We assume that OPV infections have the same duration and contagiousness whether they arose from direct vaccination or transmission from someone with an OPV infection. To further reduce parameters, we assume that ratios affecting relationships between first and reinfections are the same for OPV and WPV infections and that the ratios affecting duration and contagiousness of OPV compared to WPV infections are the same for first infections and reinfections.

4.3.3 Age structure

To examine vaccination effects, we add age to the model because vaccination is only given to children less than 5 years old. Instead of making chronological time and aging time move across the same small time steps, we reduce the computations needed for model analysis by moving age progression across larger steps (1.5 months) than chronological time and keeping track of age only until age 5. We confirmed that this does not cause inappropriate leakage of susceptibles out of the vaccine ages to older ages.

4.3.4 Modeling vaccination patterns over time

Patterns of vaccination vary greatly in the real world and extensive modeling work has demonstrated how important capturing the past pattern of vaccinations accurately is to making valid predictions of current and future risks that arise from inadequate vaccination programs in the past[28, 30, 31, 32, 33, 34, 35, 98, 116, 117, 118, 119, 120, 121, 122, 123]. But we want to show why these models will not

adequately capture the risk of prolonged SC. So, we use a generic vaccination pattern that can do that. Vaccination patterns over time were simplified by assuming all children were vaccinated at the same rate at any moment in time. Rather than fitting vaccination patterns to any one country, we characterized a general pattern found in the places where polio held on the longest. That pattern involved periods of more intense vaccination that knocked down transmission followed by less intensification of efforts or even diminution of efforts after initial success. This is followed by boosted vaccination levels that finally eliminate polio cases. The characteristic of the real world that our model of vaccination patterns sought to capture was variability in the time that WPV transmission was knocked down without being knocked out before a final boost in vaccination levels leads to elimination of polio cases. We used a linear increase in vaccination during the delay time before the final boost. When there is a delay, we begin at zero and linearly ramp up vaccinations across the delay interval to a level that achieves a first infection prevalence of 300 at the end of the ramp. That vaccination level varies by the length of vaccine ramp-up, the type of waning, and the relative transmissibility of the vaccine. It is determined by fitting the vaccination level to the prevalence level at the end of the vaccination ramp up. Using this tactic allows us to vary the time that vaccination knocks transmission down without knocking it out while keeping conditions at the time of a boost in vaccination constant. Examples of vaccination patterns are found in Figures 4.5, 4.6, and 4.7. For model settings where there is no delay, the rate of vaccination is constant at its final level from the onset of vaccination.

4.3.5 Waning model rationale and description

Given the complexity of immune control of infection, waning immunity potentially has many dimensions. We choose to reduce this complexity to a single flow

from the recovered stage (R) to the partially susceptible stage (P). Two parameters determine the waning process: the waning rate, and the waning depth or the fraction of transmission potential (basic reproduction number) that a population of all individuals in the P state would have compared to a population of individuals all in S. The waning depth has three components that we set to equal values. These are the relative susceptibility of individuals in the P compartment with respect to those in the S compartment, the relative contagiousness per unit time of reinfections with respect to first infections, and the relative duration of reinfections with respect to first infections. Because transmissibility is the product of contagiousness and duration, transmissibility wanes more than susceptibility. The waning depth is the product of three terms: the fraction of waning in susceptibility, the fraction of waning of contagiousness, and the fraction of waning of duration.

To demonstrate the effect of different waning patterns, we chose three illustrative values of the waning rate. We then found three corresponding values for the waning depth which, in the absence of vaccination, generate the same exponential distribution of ages of infection in each pair of waning rate and depth parameters. We used pair values that give a slow-deep waning pattern, a fast-shallow waning pattern and an intermediate waning pattern as illustrated in Figure 4.2 and specified in Table 4.1. Note that the quantity shown in Figure 4.2 is the product of four elements (1) the fraction of the population that would be in P state if everyone started in the R state and never experienced any subsequent infection, (2) the susceptibility of someone in the P state divided by that of someone in the S state, (3) the contagiousness of someone in the P state who gets infected divided by that of someone who got infected from the S state, and (4) the duration of infection for infections from the P state divided by duration of infections from the S state. All four of these elements

determine effective reproduction numbers as shown in Chapter V. In Section 5.3, we analyze at equilibrium how model parameters affect the ratio between transmissions from first infections and reinfections.

The average age of infection in our model is determined jointly by the basic reproduction number and by waning parameters since the first generates a force of infection from first infections and the later generates a force of infection from reinfections. We examined two different average ages of infection at equilibrium before vaccination was begun: Five years for lower transmission and 3.333 years for higher transmission. The age distribution at time of first infection for all of our high transmission settings result in an R_0 estimate of 15 using the methods of Fine and Carneiro[43]; our low transmission settings would likewise all result in an R_0 estimate of 10. These two values represent high and low values for "poor developing countries" in [43] Figure 1. The basic R_0 in Table 4.1 is the number of infections that a single first time infected individual would generate if all other individuals in the population were in the completely susceptible S state. How we set parameter values is in the legend of Table 4.1. That the values in Table 4.1 generate dramatically less waning than the model by the Institute for Disease Modeling[12, 130] is illustrated in Figure 5.7.

Our primary outcome variables of interest are duration of silent circulation and outcome of silent circulation. Silent circulation duration is defined in the model output from the time that cumulative incidence of first infections over the past year falls below the first infection to paralysis ratio (IPR) until either the cumulative incidence of first infection rises once again above the IPR (recurrence) or the prevalence falls below an average of one individual in the population (eradication). Stable eradication at a given level of vaccination occurs when the equilibrium effective reproduction

Waning scenarios in Figure 4.2	Panel	Avg. age 1st Inf no vaccine	Waning rate	Waning depth	Basic R_0	Vaccine rate at end of 20 yr ramp up	Effective contact rate
No waning higher transmission	A	3.33	0	–	15	0.25	195.3
Fast shallow higher transmission	B	3.33	0.2	0.073	9.3	0.33	121.5
Intermediate higher transmission	C	3.33	0.04	0.26	9.3	0.42	121.5
Slow deep higher transmission	D	3.33	0.02	0.6	9.3	0.45	121.5
No waning lower transmission	E	5	0	–	10	0.20	130.2
Fast shallow lower transmission	F	5	0.2	0.073	6.9	0.25	89.8
Intermediate lower transmission	G	5	0.04	0.26	6.5	0.31	83.4
Slow deep lower transmission	H	5	0.02	0.6	6.4	0.35	83.3

Table 4.1: Transmission, vaccination, and waning parameters. This analysis is performed using type 3 polio parameters for the paralysis to infection ratio of 1000 and the relative transmissibility of OPV to WPV of 0.37. The first three columns of numbers are set parameter values with the caveat that the waning depths for B–D were selected to give approximately the same R_0 values. Then those three values were used for waning depths F–H. The basic reproduction number is set to the value found to give the desired average age at first infection at equilibrium before vaccination is begun. The effective contact rates at these settings are indicated in the last column. The vaccination rates are those that give a prevalence of first infections of 300 per million at the end of the ramp up period and are found separately for each of the eight settings given the values in the first four columns.

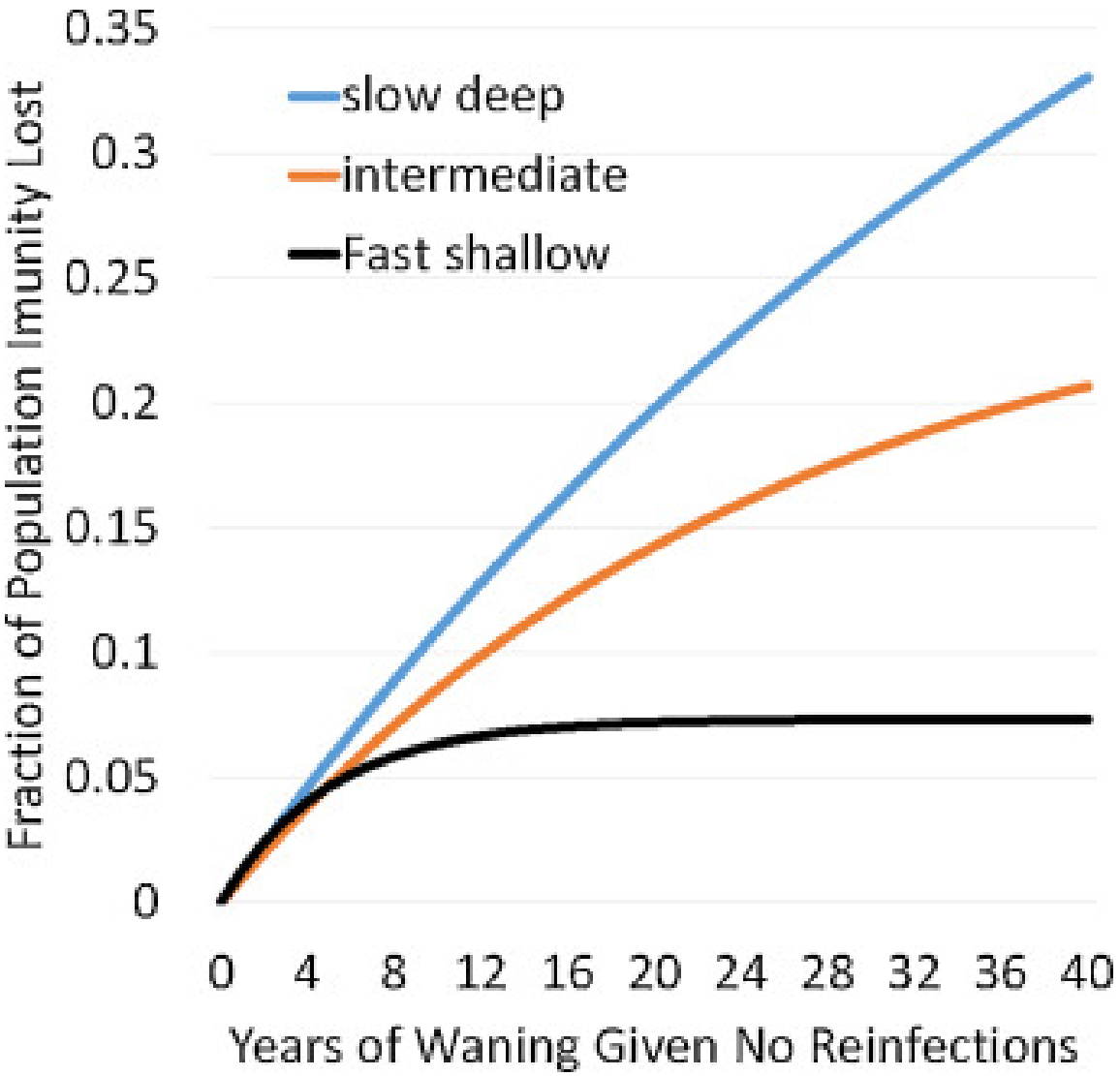


Figure 4.2: Population extent of waning for the different waning scenarios as defined by waning depth affecting susceptibility times waning depth affecting duration, times waning depth affecting contagiousness under the assumption that all three effects have the same waning depth parameters. Waning rate and depth parameters used are from Table 4.1.

number from both first infections and reinfections is below one. The reproduction number may continue to rise for many years after the last polio case when there is slow or intermediate waning. Stability, therefore, is determined by running models for a long time. Given stable eradication, reintroduction of infection from the outside after local transmission has ceased will never result in an epidemic. With unstable eradication, however, individuals with waned immunity in the P state from Figure 4.1 will eventually put the effective reproduction number above one. True global eradication that eliminates all virus in people and laboratories would be practically stable but could be mathematically unstable.

4.3.6 Model equations and numerical solutions

The model equations are presented in Chapter V. The code used to implement those equations in Berkeley Madonna (Oster and Macey, 2015) can be obtained by the means discussed in Appendix B.

4.3.7 Model analysis

The model is analyzed both mathematically and numerically. Numerically the model is run to WPV equilibrium before introducing OPV vaccination. Numerical analyses were checked to see that they were not affected by the numerical integration algorithms used or the time steps employed. Mathematical analysis is presented in Chapter V.

4.4 Results

4.4.1 Model generated patterns of drop in first infection prevalence

Figure 4.3 shows the pattern of decreasing incidence of first infections during a 20 year ramp up delay of vaccination for parameters in Table 4.1. Given the infection to paralysis ratio of 1000 used in this analysis and the prevalence of first infections

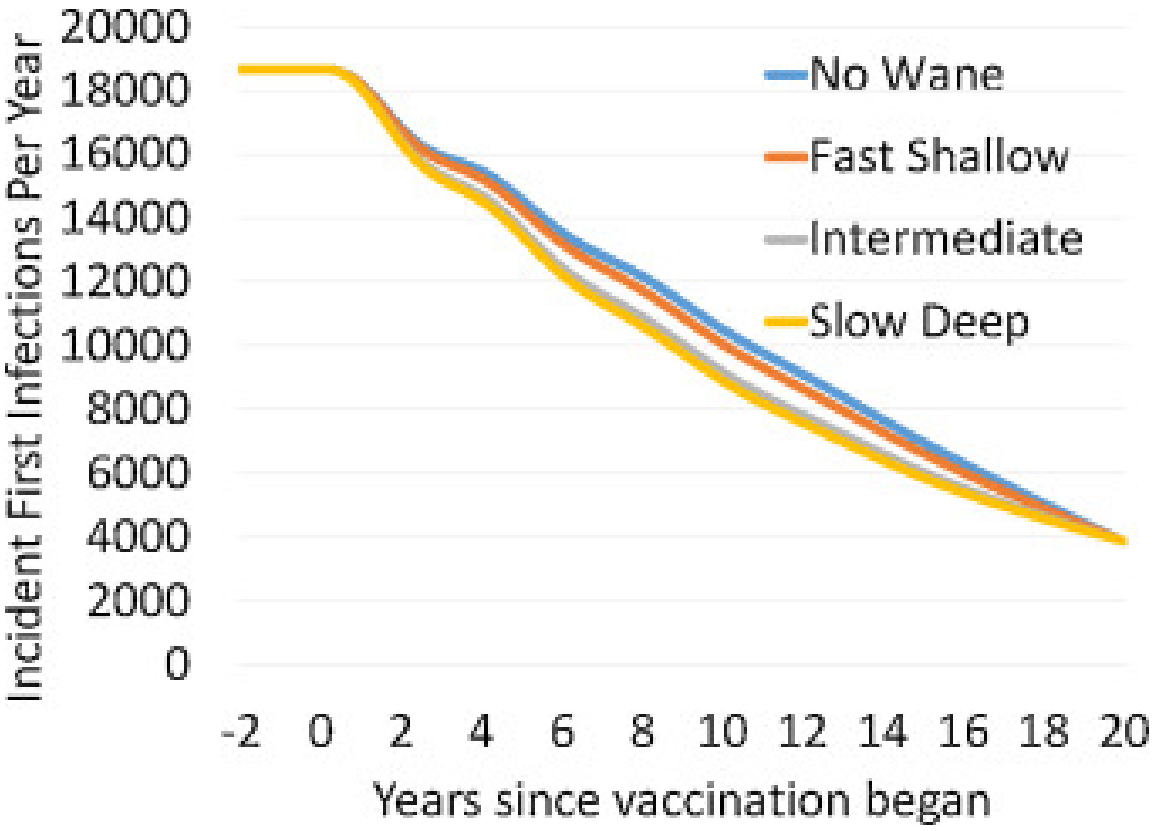


Figure 4.3: First infection incidence as a function of waning scenario across a 20 year delay period during which first infection prevalence is reduced to 300 per million for the higher level of transmission. Settings used are those for a type 3 virus and settings A–D from Table 4.1.

of 300 at the end of the ramp up, 3–4 polio cases per year in our population of one million would be expected at the end of the ramp up.

The procedure for setting parameters outlined in the legend of Table 4.1 was chosen to make the force of infection identical at both the endemic equilibrium and at the end of the ramp up, for all waning scenarios. By doing this, we demonstrate that the pattern of polio cases over time in a population cannot be used to determine the waning pattern in that population. That inference is important because in Section 4.4.2 we will show that waning patterns with similar effects on the population level of infection before vaccination is begun have very different effects on silent circulation duration after the last polio case. Note that there is a tradeoff between settings

for waning parameters and Basic R_0 (or the contact rate used to set that) and that we chose values for the set waning rates and depths that kept the average age at first infection constant for the higher transmission settings given that there was any waning at all. That multi-parameter tradeoff makes the lack of identifiability illustrated in Figure 4.3 even greater. The way we present Figure 4.3 emphasizes the role of effective vaccination rates as a single factor affecting identifiability. If effective contact rates, average age of first infection in the pre-vaccine era, and infection to paralysis parameters were known, adding information about effective vaccination levels might theoretically help determine waning patterns. From Table 4.1 we see there is some increase in vaccination level needed to get to our fixed prevalence of first infections as we progress through fast-shallow to slow deep waning scenarios. Data and modeling of vaccination rates has greatly increased in recent years[98]. But given the difficulties in determining what the effective levels of vaccination have been in populations over longer times, the waning patterns will not be identifiable from standard polio program data.

4.4.2 Silent circulation patterns

SC begins in our model when vaccination brings down cumulative incidence of first infections over the past year to less than the first infection to paralysis ratio. That point is a likely time when the last polio case might be detected. SC ends with either mathematically stable eradication, mathematically unstable eradication, or recurrence of polio cases. We declare eradication when the prevalence of infection goes below one. If the cumulative incidence of first infections since the last polio case rises above the IPR before the prevalence goes below one, then we declare a recurrence of polio cases.

Figure 4.4 presents SC duration by the final stable rate of effective vaccination

of children under age 5. This is lower than the actual vaccination rate. It is the rate at which vaccine infection occurred and stimulated immunity in completely susceptible children. Two patterns of vaccination are presented. In the first there is no delay in vaccination reaching its final levels. In the second there is a 20-year delay during which vaccination is linearly ramped up to a level that generates a first infection prevalence of 300 per million total population. The prevalence paths taken to this value are indicated in Figure 4.3. Parameters used are indicated in Table 4.1. Whether the SC ended with a recurrent polio case, unstable eradication where a new introduction could lead to a new outbreak, or stable eradication where outbreaks would be small is indicated by color for different vaccination levels. The no delay curves always occur at lower vaccination levels than the 20-year delay curves.

To help understand where the points in Figure 4.4 come from, let us look at Figures 4.5–4.7. Then we will return to discuss what the patterns in Figure 4.4 indicate. Rather than presenting levels of infection, Figures 4.5–4.7 present the contributions of first infections and reinfections in the populations under or over age five to the total effective reproduction number. We made our model simple enough so that the contributions of each group of infections sum to the total effective reproduction number. So let us first explain those contributions.

Figure 4.4 reveals 3 effects as waning progresses from no waning to slow deep waning across panels A–D and E–H. As waning becomes slower and deeper while keeping the effects of waning on age at first infection constant before vaccination begins, (1) it takes higher final vaccination rates for eradication to be achieved. (2) The vaccination rates required for eradication increase between situations where final vaccination rates are instantly achieved or have a 20-year delay. (3) The breadth of vaccination levels that result in prolonged silent circulation increases.

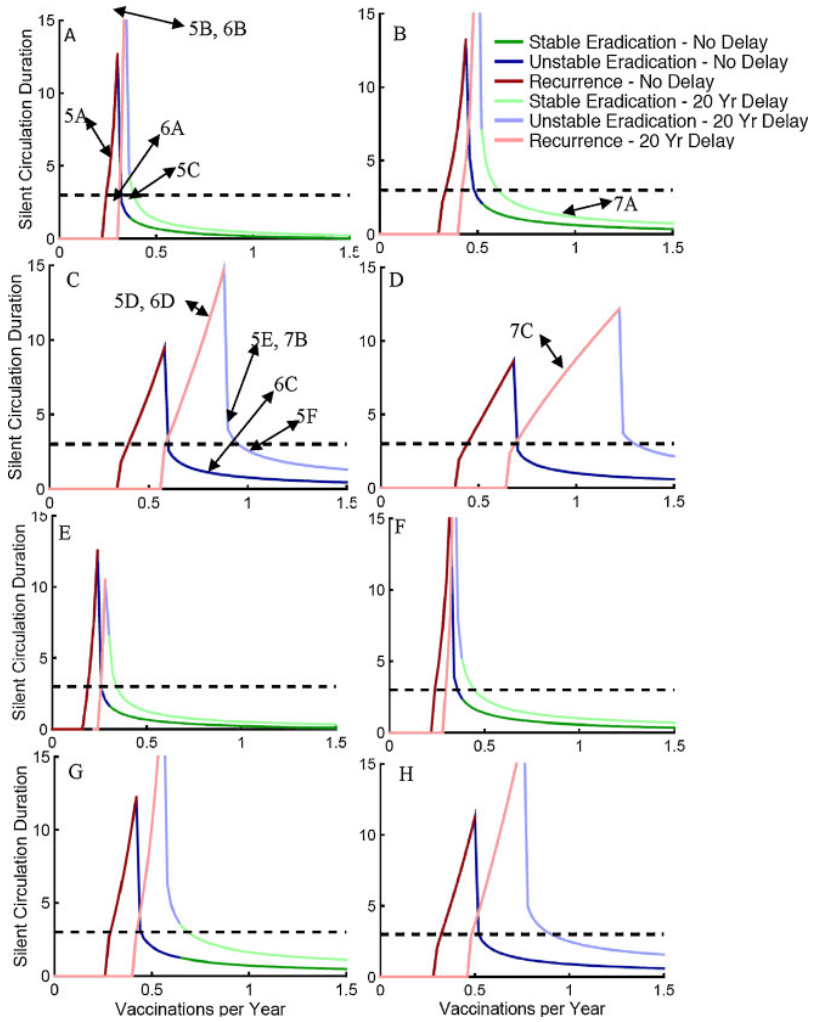


Figure 4.4: Patterns of duration of silent circulation as a function of final total vaccination rates in less than five year olds by waning immunity scenarios: (A and E) no waning, (B and F) fast shallow waning; (C and G) intermediate waning; and (D and H) slow deep waning. (A–D) have an average age of first infection of 3.33 years while E–H have an average age of 5. All panels have type 3 settings (IPR = 1000, OPV/WPV transmission ratio = 0.25). Type of silent circulation (stable, unstable, and recurrent) is indicated by color. Horizontal dashed line illustrates 3-year silent circulation. Parameter values used are shown in Table 4.1. The points on these curves where the dynamics are illustrated in subsequent figures are indicated by figure number and panel letter of Figures 4.5–4.7.

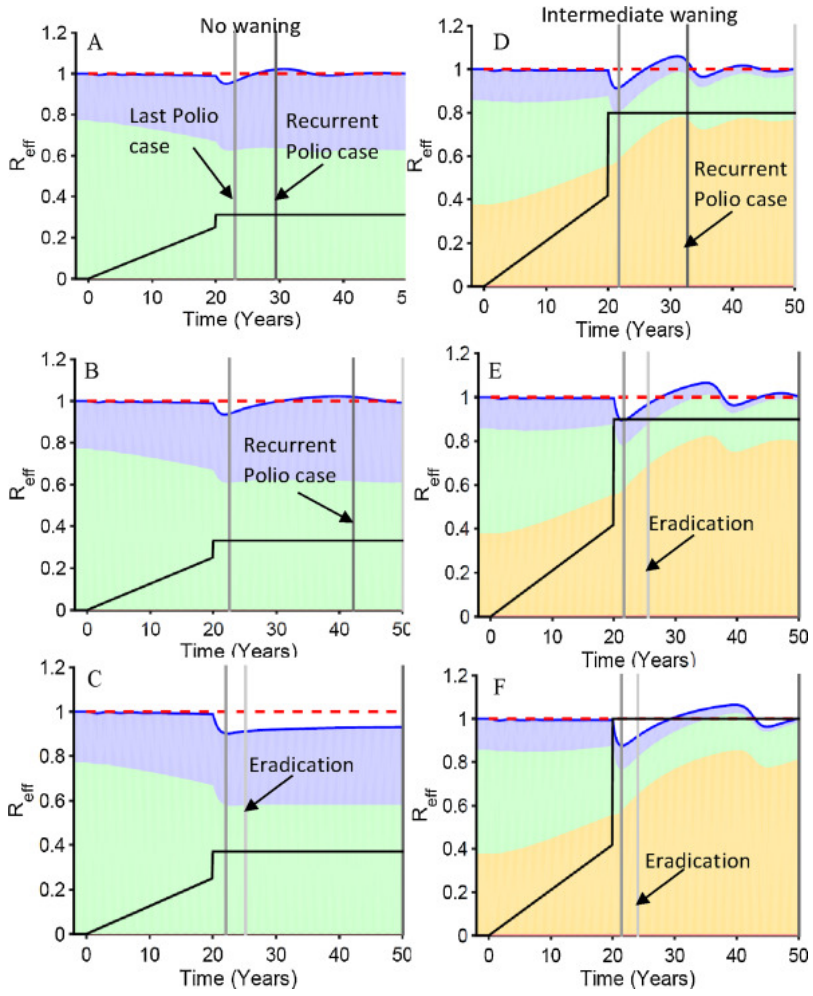


Figure 4.5: The dynamics of the effective reproduction number and its components as vaccination levels are increased. Intermediate waning, the higher level of transmission, a 20-year delay and serotype 3 settings were used. The position of each panel in Figure 4.4 curves are indicated in Figure 4.4. The final vaccination levels for panels are A = 0.32 effective vaccinations per year, B = 0.34, C = 0.38, D = 0.8, E = 0.9, F = 1.0. The black vaccination level lines ramping up to 20 years and then jumping to the final vaccination level use the same left hand scale as the effective reproduction numbers. The blue shaded area is the contribution to the effective reproduction number from first infections in the five and older age group, the green is first infections from the under 5, the yellow is reinfections in the five and older, and the red (barely visible below the yellow) is from reinfections in under age five. The dark blue line summing all shaded areas is the effective reproduction number. The red dotted line is the transmission threshold. The time of eradication, last polio case, and a recurrent polio case have increasingly darker gray shading.

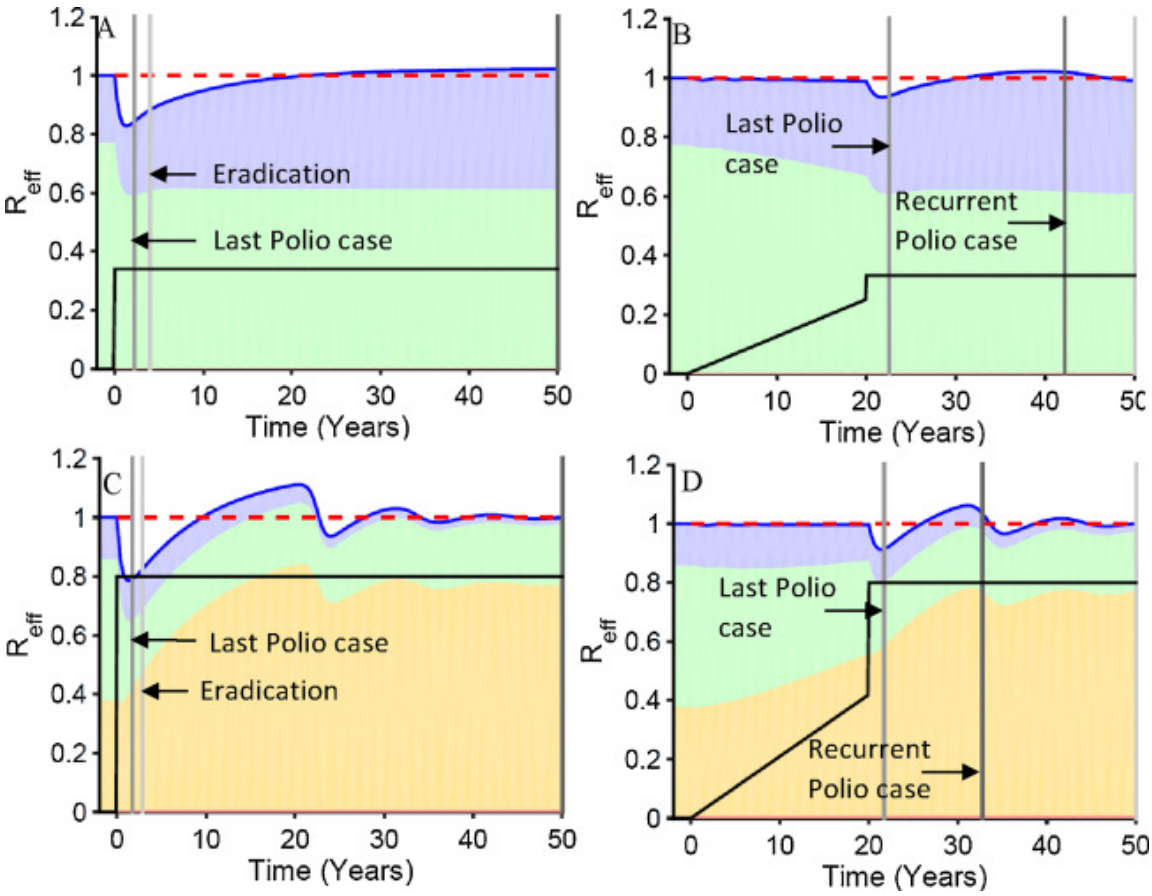


Figure 4.6: The dynamics of the effective reproduction number and its components with and without a delay in reaching final vaccination levels. Intermediate waning, the higher level of transmission, and serotype 3 settings were used. The positions of each panel on silent circulation duration curves are indicated in Figure 4.4. The shading and lines have the same meaning as Figure 4.5.

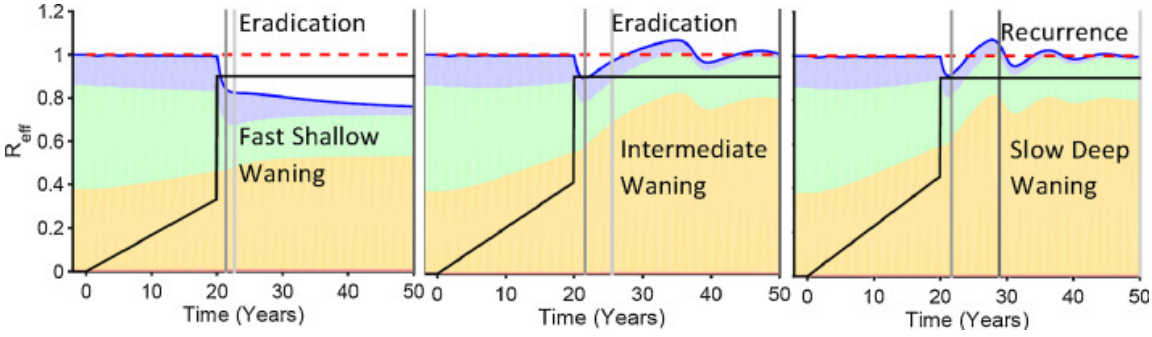


Figure 4.7: Changing dynamics of the elements of the total effective reproduction number across different waning scenarios explain the effects of those waning scenarios when final total vaccination levels are kept constant. Higher transmission parameters along with IPR and OPV/WPV transmissibility characteristic of type 3 were used.

These differences are greater when transmissibility is at the upper end of the plausible range for higher transmission countries (A–D) than when it is at the lower end of that range (E–H). The differences between no waning and fast shallow waning are small. In these scenarios, there is no meaningful ongoing waning after 5 years, so that there is no ongoing accumulation of people susceptible to reinfection after 5 years. As seen in Figure 4.2, there is such an ongoing accumulation of partially susceptibles with intermediate waning, and there is twice as much accumulation with slow-deep waning.

4.4.3 Effective reproduction number analyses to explain silent circulation (SC) patterns

The total effective reproduction number for our model is the sum of the effective reproduction number for first infections plus the effective reproduction number for reinfections in each of the age groups. The mathematical logic supporting this is shown in Section 5.2.

$$R_{\text{EffTot}} = R_{\text{EffFirst}<5} + R_{\text{EffSubsequent}<5} + R_{\text{EffFirst}\geq 5} + R_{\text{EffSubsequent}\geq 5}$$

When R_{EffTot} is less than one, infection levels will be falling. When it is greater than one, they will be rising. The dynamic pattern of these four components helps determine whether a given vaccination level will result in stable eradication, unstable eradication, or a recurrent case of polio. How far vaccination drives the total effective reproduction number below one and how long it stays there determines how far infection levels will drop. The effective reproduction number when vaccination is first begun always starts out at the value of one in all the dynamics panels in Figures 5–7 because we start vaccination only when the system is at equilibrium.

4.4.4 Temporal dynamics leading to the summary points in Figure 4.4

Figures 4.5–4.7 color the four contributions to R_{EffTot}. Green is the contribution from first infections in the under-five age group. Blue is first infections in the over-five group. Yellow is the reinfections in the over-five age group. Red is reinfections in the under-five age group. The contributions of this latter group are so small that they are not readily perceptible in the figures. The time when the last case is seen is indicated by the earliest vertical line. The time after that when a recurrent case is detected (when the sum of new first infections since the last case exceeds the IPR) is indicated by a darker vertical line. The time when eradication occurs (the sum of first infection in the past year becomes less than the IPR) is indicated by a lighter vertical line. The temporal patterns of vaccination levels are indicated on these figures using the same scale as the effective reproduction number.

We do not stop all transmission in these figures when the number of total infections, including first infections and reinfections, goes below one and we declare eradication. Of course, once there is eradication, in the absence of reintroductions the effective reproduction goes to zero because there are no infections. But for didactic reasons we follow dynamics beyond this point by allowing a fraction of an individual to continue transmitting. This helps us see whether eradication is stable or unstable. If the effective reproduction number ever goes back above one after eradication, the eradication is unstable and introduction of a new case could set off a large epidemic.

Note that in Figure 4.5 before and during the 20 year ramp up, all panels are identical within the two waning scenarios examined (A–C no waning and D–F intermediate waning). During the 20 years of vaccination ramp up, the total effective reproduction number dips only slightly below one. But it dips enough down to drive

the endemic prevalence of first infections to 300. For the no waning scenario in panels A–C of Figure 4.5 there are only first infection events so the above equation has only two elements rather than four because with no waning, there are no reinfections. During the ramp up delay, vaccination reduces the contribution of first infections in the under-five age group. But that drop is compensated for by more children being susceptible at their fifth birthday so that the greater than five-year-old susceptible group contributes more to the effective reproduction number. When there is no waning, it takes only a small boost in vaccination levels to lower the effective reproduction number to where cases are no longer detected (Panel A). A very slight increase in the final vaccination level can extend the SC interval between the last case and the time when a recurrent case is detected (Panel B). Then a further small increase (Panel C) gets the effective reproduction number continuously below one so that eradication is stable and only small outbreaks would occur if there were a new introduction.

Panels D–F show a similar pattern of the total effective reproduction number staying near one during the ramp up for intermediate waning. But given waning, the forces driving R_{EffTot} back up toward one now mainly arise from reinfections in individuals with waned immunity. These reinfections replace first infections in both the groups under the vaccination age and over it. The decrease in the force of infection in the vaccination age group means that vaccination levels must be higher to get the same decrease in the effective reproduction number.

A final thing to notice in the intermediate waning panels of Figure 4.5 is that once eradication is achieved, further waning raises the contribution of reinfections so that it only takes very low levels of susceptibility in the under-five age group to get the effective reproduction number above one. Thus, eradication in the presence of

waning is unstable.

In Figure 4.6 we see how a delay affects the dynamics in the absence or presence of waning. Both in the presence and in the absence of waning, a delay causes a decrease in the size of the drop in the total effective reproduction number when vaccination is boosted. In both cases, that decrease occurs because the contribution to the effective reproduction number by the age group getting vaccinated decreases and is replaced by contributions from other groups. When there is no waning, it is the contribution from the older age group who did not get vaccinated that increases during the delay. With ongoing waning, it is primarily the contribution from older individuals whose immunity has waned that increases. Again, we see here that when there is ongoing waning, it takes much higher levels of vaccination to get elimination of polio cases and eradication of transmission.

Note that the difference in the compensating groups implies different policies for addressing a failure to eliminate polio cases or a recurrence of polio cases after their elimination. If there is no waning, then the increased susceptible individuals over age 5 will all have accumulated during the delay and thus young people should be targeted for special control programs. But in the presence of ongoing waning, it is the age groups that have passed the most time since vaccination or natural infection whose immunity is most likely to have waned. Thus, older, rather than younger, age groups should be targeted.

Figure 4.7 shows how waning patterns affect the outcome of eradication efforts given nearly identical patterns of decrease in polio cases as shown in Figure 4.3. All three panels have waning that produces the same average age of first infection before vaccination is begun given the same level of transmissibility as shown in Table 4.1. We focus on three times on the horizontal axis of each panel: (1) before vaccination,

(2) during the ramp up, (3) after the boost.

Before vaccination is begun, all three waning scenarios have the same fractions of age and immunity groups that contribute to the total endemic effective reproduction number of one. During the ramp up, as the depth of waning increases it takes higher vaccination levels to get to a first infection prevalence of 300. The largest jump is between fast-shallow and intermediate waning. Likewise, there is a considerable difference between fast shallow and either intermediate or slow deep waning regarding the buildup of the fraction of the effective reproduction number that comes from reinfected individuals. Fast shallow waning levels off so that there is no ongoing waning after the first few years but any ongoing waning, whether it is a little or a lot, makes a big difference in dynamics.

Finally, we look at the response to the boost given the same final level of vaccination. The difference between those scenarios with and without ongoing waning is dramatic. There is very little rise in the contribution of reinfections for the fast-shallow scenario and consequently the effective reproduction number drops dramatically and eradication is achieved rapidly and stably. In contrast, the rise in the contribution of reinfections is higher in the intermediate and slow-deep scenarios, and in both, eradication is achieved slowly if at all, and unstably. The difference between intermediate waning and slow deep waning is enough to shift the end of SC from mathematically unstable eradication to recurrence of detected paralytic cases.

4.4.5 Patterns of SC with increasing boosts to higher final vaccination levels with and without delays

Examining Figures 5–7 and seeing where each of the specific cases in these figures fall on the curves in Figure 4.4 should now give the reader a better understanding of what Figure 4.4 represents. Note in Figure 4.4 that the range of vaccination levels

that result in prolonged silent circulation increases with increased waning across panels A–D and E–H. It also increases in each panel between the no delay case and the delay case with a 20-year ramp up to levels of vaccination that still allow for detectable paralytic cases. Finally, it also increases with increased transmission as seen by comparing A and E, B and F, C and G, and D and H. There is little difference between no waning and fast-shallow waning scenarios but bigger differences between fast-shallow waning and intermediate level waning. Those differences are augmented given vaccination delays and greater transmission.

A big determinant of how likely we are to encounter prolonged silent circulation in the real world is the width of the vaccination levels across which there is prolonged silent circulation. If vaccination levels must be held in a narrow range for there to be prolonged silent circulation, then natural fluctuations and chance effects will eliminate that risk. Thus, one conclusion from Figure 4.4 is that prolonged silent circulation will become increasingly likely in the real world as the delay in achieving effective vaccination levels increases and as the amount of ongoing waning increases. These two factors are more important than the level of transmissibility in determining silent circulation risks. The effects of these factors, however, are greater at higher levels of transmissibility.

This is important because empirical experience with prolonged silent circulation risks is based on experience in settings with lower transmission levels and shorter times between beginning vaccination efforts and ending polio cases than are characteristic of the currently or recently endemic countries. Thus, there is reason to expect that empirical experience may not be a good guide for setting the time period after the last polio case at which we are confident that there is no ongoing silent circulation.

Another thing to note is that higher vaccination rates are required to eliminate paralytic cases as the depth of waning is increased. This is noted by seeing where the curves arise above the zero level. The effects of delays in reaching needed vaccination levels and of increasing transmission on the levels of vaccination needed to eliminate detections of paralytic cases are analogous to their effects on increasing the range of vaccination levels with prolonged silent circulation.

Another observation is that, in our model, prolonged silent circulation almost always ends with a recurrent case. Once a high enough level of vaccination is reached to eradicate infection, there is only a narrow range of vaccination levels where it takes more than three years of SC before eradication levels are reached. As final vaccination levels in our model are raised, the first eventful occurrence is that more than a year passes without a paralytic case so that our curves rise above the zero level. Then it takes only a tiny further increase in vaccination levels to get the duration of SC above 3 years. With higher vaccination levels, the time of a recurrent case is pushed further into the future so the duration of SC increases. But the duration of SC is limited by finally reaching a level where eradication occurs. The peaks in Figure 4.4 curves always represent scenarios where SC ends with a recurrent case.

A final observation is that when eradication occurs, whether it is stable eradication where new introductions only cause small outbreaks, or unstable eradication where long chains of transmission can be sustained, depends strongly on the depth of waning. Eradication in the no waning and fast-shallow waning scenarios is mostly stable eradication while stable eradication is almost never seen with intermediate or slow-deep waning. Without ongoing waning after a few years, there is not enough of a buildup of partially susceptible individuals with waned immunity after eradication occurs to get the total effective reproduction number above one should there be a

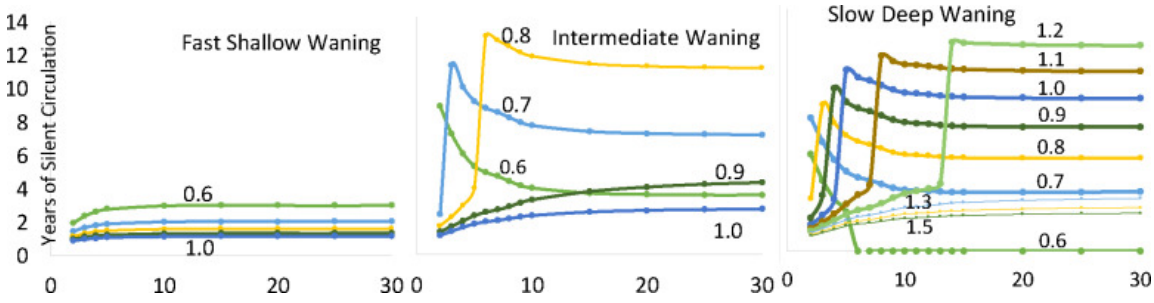


Figure 4.8: Years of silent circulation (vertical axis) by type of waning (in different panels), years between beginning a vaccination program and lowering first infection prevalence to 300 per million, (horizontal axis) and the rate per year to which vaccination of children under 5 is boosted when the prevalence of first infection reaches 300 (labels on each line). The horizontal axis is plotted only at a 2-year delay and above.

reintroduced case.

4.4.6 A different graph for the same phenomenon

Instead of graphing the final vaccination level in a continuous manner and the delay in time as discrete, it is instructive to switch to make the delay more continuous and the final vaccination level more discrete. We do this in Figure 4.8. For that figure, for each delay time we calculate the vaccination rate at the end of the delay period to give a prevalence of first infection of 300 individuals in our population of one million.

In Figure 4.8, the effect of increasing waning depth on the range of vaccination levels across which there is prolonged silent circulation is seen just as it is in Figure 4.4. In fast-shallow waning there is no prolonged silent circulation at a vaccination rate in children under five above an effective vaccination level of 0.6 per child per year. In intermediate waning, there is prolonged silent circulation from 0.6 to above 0.8. And in slow deep waning, there is prolonged silent circulation from 0.6 to above 1.2.

This graphic, however, shows something not seen in Figure 4.4. It uncovers in more detail how the length of the delay in finally achieving effective vaccination

levels interacts with both waning depth and final vaccination levels to affect the duration of SC. As delay time increases in the presence of ongoing waning beyond the first few years (intermediate and slow-deep waning in this figure), there is a sharp transition from a specific vaccination level resulting in relatively short SC that ends in eradication to much longer SC that ends in a recurrent case. As the delay time increases, both the vaccination levels that undergo this transition increase and the duration of the SC increase. At some vaccination levels after a boost, there appears to be a threshold for the length of the delay where reinfections cause prolonged low-level SC. This threshold, although generated by reinfections, is different from the reinfection threshold discussed by[49]. The threshold occurs as reinfection is increased, as in [49]. But here the threshold is for SC, not any circulation.

4.4.7 Serotype differences in silent circulation potential

Our model assumes that serotypes differ only in their first infections to paralytic infections ratio (IPR) and in their OPV transmissibility to WPV transmissibility ratio or vaccine transmissibility which we label as VT. We use values of VT of 0.37, 0.57, and 0.25 and values of IPR of 200, 2000, and 1000 for types 1, 2, and 3 respectively. These values are in the range used by other modelers[31]. In this model, Serotype 3 is the hardest to eradicate and Serotype 2 the easiest (Figure 4.9). The risks are smaller for the no waning scenario and the fast-shallow waning scenario and the changes in silent circulation for these first two waning scenarios are mostly in the level of vaccination where prolonged silent circulation is seen rather than the width of the vaccination range across which it occurs. But for intermediate and slow-deep waning, the effects on the width of the vaccination interval resulting in prolonged silent circulation is quite large.

To separate the effects of the IPR and VT we compared settings starting with the

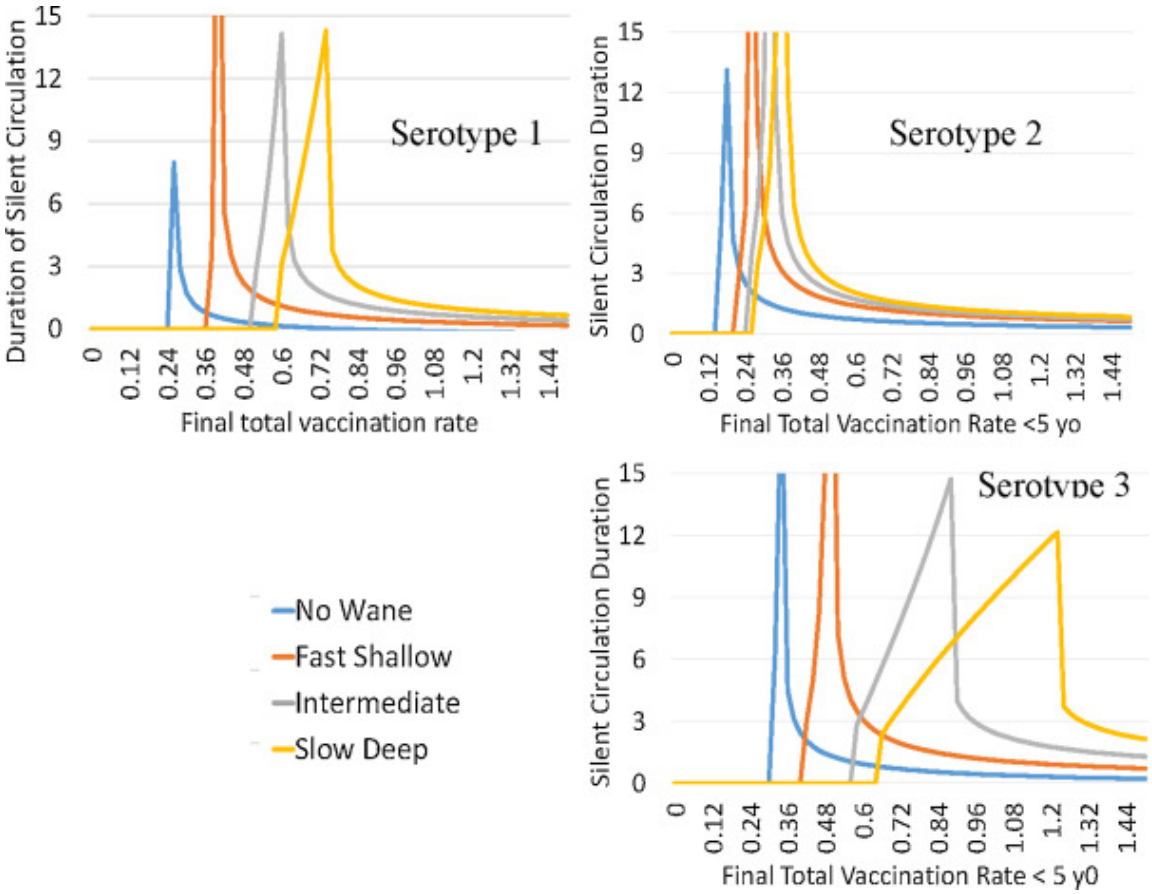


Figure 4.9: Length of silent circulation by final total vaccination rates in less than five year olds and waning scenarios for different serotypes. Settings include a 20 year ramp up time delay and high transmission.

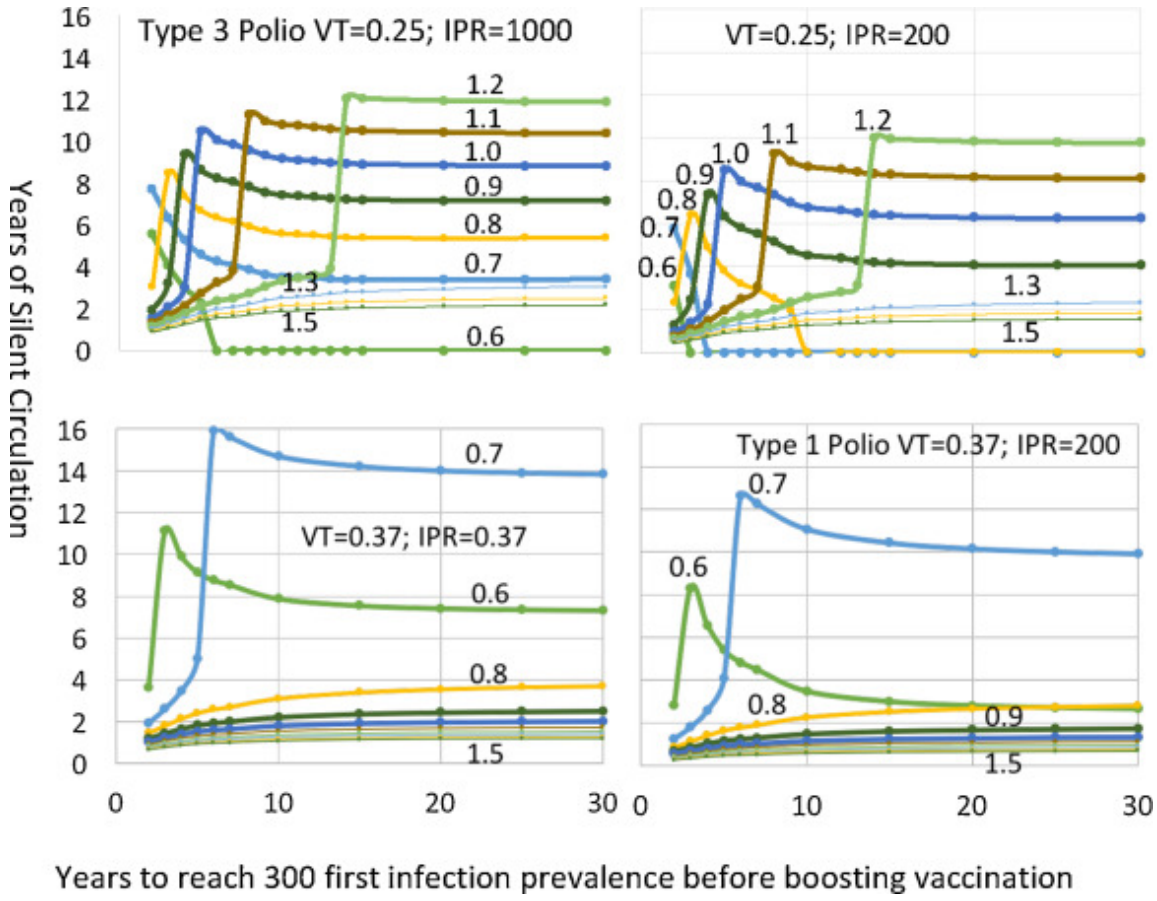


Figure 4.10: Years of silent circulation for slow deep waning by the ratio of OPV transmissibility over WPV transmissibility, infection to paralysis ratio, and the number of years to ramp up vaccination just enough to reach a prevalence of 300 first infections before boosting vaccination to the final levels indicated by the labels above the lines.

serotype 3 settings of 1000 and 0.25 and changed them to 1000 and 0.37 and 200 and 0.25 so that we could see more clearly what leads to the difference between serotype 1 and serotype 3. The graphs with continuous delay times and discrete vaccination levels show very clearly the different effects of the two parameters varied. In Figure 4.10 we see that the big effect on the risk of SC is the lower vaccine transmissibility (VT) rather than the higher IPR. A lower IPR lowers the duration of SC given that there will be a recurrent case because it makes it easier to detect the recurrent case. Inadequate AFPS would have the same effect as a high IPR. But a lower VT leads to long duration SC at higher vaccination rates.

This graph also shows how the two causes of SC (poor AFPS and reinfection transmission) interact. In our model, lowering VT increases reinfections because less vaccine transmission means less boosting of waned immunity. Raising the IPR could be a result of poor AFPS. In Figure 4.10 we see that poor AFPS slightly lengthens any silent circulation generated by the mechanisms we have illustrated in our analysis. But increasing waning by decreasing vaccine boosting has two effects. First, for any given vaccination boost amount after a delay in eliminating cases, if the SC threshold is passed, SC duration levels out at a shorter value. Compare the 0.7 vaccination boost level to see that. That is because more waning shortens the time to a recurrent case. But the second effect could be more important. More waning means that the silent circulation threshold can be passed with much higher vaccination boosts. That is because vaccination boosts affect the completely susceptible children being vaccinated less because reinfections have displaced them in the overall effective reproduction number near one before the boost in vaccination. Consequently, as illustrated in Figures 4.5– 4.7, the effective reproduction is dropped less. It is also dropped for less time because there are more individuals with waned immunity to bring the effective reproduction number back up toward one.

4.4.8 Levels of prolonged silent circulation

Prolonged silent circulation that eventually results in eradication in the situations we examined had a maximum prevalence of wild polio virus infections summing across first and repeat infections of 110 per million or lower for the slow deep waning and 81 for the intermediate waning. Those levels were highest at the time of the last polio case and decreased progressively with time unless there was a rebound leading to a recurrence of infection before there was eradication. For first infections those top values are 50 and 32. It could thus take a considerable time to see a paralytic

polio case and missing a polio case in AFPS surveillance could double that time.

For the scenarios that ended in a recurrent case, the maximum prevalence occurred when the recurrent case appeared right after three years. For both slow deep and intermediate waning, the prevalence was 200 per million at the time of the last polio case. It went down for a year, and then came back just before three years to 500.

These levels would take extensive environmental surveillance (ES) to detect. Not only would ES need to detect a single reinfection in every population of 10,000, but the detection sensitivity would have to be more than 10 times greater than that needed to detect a first infection. Thus, ES as currently being implemented is likely to miss the prolonged low level silent circulation that arises in our model.

4.4.9 Levels of vaccination in older age groups needed to avoid prolonged silent circulation.

To evaluate the effect of vaccinating older age groups before OPV use is stopped, we determined the rate at which the five and older age group would have to be vaccinated to bring about eradication in less than three years from the last polio case. The worst-case from Figure 4.4 is the high transmission scenario with slow deep waning after a 20-year delay. In this situation, only 8% of the ≥ 5 population needs to be vaccinated per year to eradicate infection within three years of the last polio case. At higher final vaccination levels of the < 5 population it takes less. If the ramp up goes to a lower prevalence of polio before the vaccination is boosted to final levels, it also takes less vaccination in the older age group to ensure that silent circulation is not prolonged.

Since the model analyzed here does not break down the ≥ 5 population into adults and older children, we cannot say what age groups are the most important to vaccinate to boost waned infections. But that would seem to be adults and not

older children since the adults will have experienced more time since getting directly vaccinated with OPV and have less exposure to children who have been recently vaccinated. When vaccination of the older age groups is only during the third year after the last polio case, it takes considerably more vaccination of the older age groups to eliminate prolonged silent circulation. In our worst-case scenario discussed above, it takes nearly 1.4 vaccinations per adult during that final year. But if vaccination of the older age groups occurred during the second and third years after the last polio case, only 29% of the older age group would need to be vaccinated per year.

4.5 Discussion

Our analyses develop theory about the final stage of polio eradication that involves stopping OPV use. The implications of that theory are that dynamics that have not been previously identified could generate prolonged low-level silent circulation (SC). We only examined a simple model that helps us develop the needed theory. The task of using that theory to better guide polio eradication decisions is left to the future.

4.5.1 Major findings

4.5.1.1 The dynamics behind SC uncovered by our model analysis

In our model, waning immunity that continues for many years(ongoing waning) leads to prolonged, low-level SC under conditions that emerge only after long delays in lowering polio levels in high transmission countries. Purely short term waning does not have this effect. We adjusted the total amount of waning so that every waning scenario we examined generated the same average age of infection in the pre-vaccine era. We show that dynamics driven by forces that bring the effective reproduction number back to a value of one lie behind this risk generation. Vaccination reduces WPV infections that boost waned immunity while increasing OPV infections to a

lesser degree. Consequently, older individuals not subject to vaccination make a greater contribution to the effective reproduction number while children in the ages receiving vaccination make a lower contribution. Therefore, vaccination lowers the effective reproduction number less in the presence of ongoing waning given many years of failure to eliminate polio. At the same time, vaccination raises the number of reinfections which are not detected by AFPS. These two phenomena increase the duration of SC over time.

Another way of thinking about these dynamics is to consider what is keeping the R_{EffTot} near one when vaccination has been ramped up slowly. When vaccination of children lowers the R_{EffTot} below one, a force driving the effective reproduction back up toward one comes from reinfections in individuals with waned immunity. This force increases with increasing levels of vaccination since fewer individuals with waned immunity will have their immunity boosted by WPV infections. That means children under age five are replaced in R_{EffTot} by older individuals with reinfections. It thus takes higher rates of vaccination to get the same lowering of R_{EffTot}. This phenomenon likely explains why such high levels of vaccination are or were required to eliminate cases in India, Nigeria, and Pakistan.

In the real world, we cannot see this effect of decreased boosting of waned immunity from increased vaccination. That is because in the absence of environmental surveillance (ES) we never detect reinfections. And even with ES, we never know if we are detecting viruses from first infections or reinfections. Another important effect of ongoing waning combined with slow ramp up to adequate vaccination levels is that it increases the gap between vaccination rates that eliminate paralytic polio cases and those that eradicate infection. This makes it more likely that SC will persist. No matter how a model is made more realistic, this effect seems likely to persist

and create conditions where the risk of prolonged low-level silent circulation could have emerged only recently in the eradication program. The remaining question then is whether the total accumulation of individuals with waned immunity is enough to sustain SC. Our model examined only homogeneous populations deterministically. In stochastic models with localized contact, spotty, local, and migratory transmission at even lower levels than found by our deterministic models is likely. In that case, local die-out of transmission will shorten the duration of SC. However, we hypothesize that the same basic dynamics we have elucidated that lead to prolonged low-level SC in our models will have similar effects in such models.

4.5.1.2 SC from these mechanisms has low infection levels

Because vaccination that can eliminate paralytic poliomyelitis greatly lowers the contribution of first infections to the effective reproduction number, the levels of infection that sustain SC in our models are low. Such low levels might require extensive environmental surveillance (ES) to be detected. But increased acute flaccid paralysis surveillance (AFPS) is less likely to detect silent circulation.

4.5.1.3 Effects of transmission potential on SC

In our model, even at the deep level of waning we examined, there is little potential for prolonged SC at levels of transmission that characterized industrialized countries before 1950 using the criteria of[43]. Within their range of poor developing countries, however, higher transmission potential augmented the effects of vaccination delays, of reduced OPV transmission, and of waning in generating SC.

4.5.1.4 Effects of oral polio vaccine (OPV) transmission and acute flacid paralysis surveillance (AFPS) on SC

We demonstrated how infection to paralysis ratios (IPR) and the relative transmission of OPV to WPV both independently affect SC (Figure 4.10). Poor AFPS effectively raises the IPR. But the effect of this on SC is modest compared to the effect of lower OPV transmission. That is because lower OPV transmission together with long times to reach adequate vaccination levels can raise the number of individuals with waned immunity to a threshold level that sustains prolonged silent circulation.

4.5.1.5 Strategies to prevent SC

We found that if vaccination of all age groups was conducted in our model during the three years between the last cases and the declaration of eradication, SC duration could be lowered considerably by vaccinating small fractions of the over-five population. Three years were needed to ensure that the effective reproduction number was kept low enough for long enough to achieve eradication. One year required vaccination levels that might be difficult to achieve.

4.5.1.6 The non-identifiability of SC risks using only AFPS data

A key finding in our study is that the potential for prolonged SC could be missed by models fitted only to polio case data. Figure 4.4 shows that the different waning scenarios that have dramatic effects on the duration of SC (as seen in Figure 4.4) can have only small differences in first infection patterns. The differences in Figure 4.4 would not be detectable given expected noise from imperfect surveillance and chance events. In other words, we show that nearly identical patterns of decreasing cases on the path to polio elimination can be consistent with models having either

low or high risks of prolonged SC. Thus, models fitted to AFPS data alone will have no practical capacity to predict how long SC will continue after the last polio case. This may explain why two recently published studies found little risk of prolonged silent circulation[41, 69].

4.5.2 Relating our analyses to other studies

4.5.2.1 Kid Risk analyses

The model in this chapter is not only simpler than other models currently being used to guide polio eradication. It is also simpler than our previous published model in [86]. We became convinced from our exploration of that model that important effects on SC risks were being missed by the detailed models developed by the Kid Risk modeling group as exemplified in [31]. To clarify the dynamics behind this, we formulated a model where the contributions of different parameters and population groups could be mathematically separated. To guide policy, the simplifying assumptions in our model need to be realistically relaxed to see if that affects specific policy decisions. A good starting point for identifying simplifying assumptions that are important to relax and others that are unlikely to make any difference is to understand population dynamics level mechanisms well. Our analysis pursues that goal as a first step in the process of inference robustness assessment[75].

The Kid Risk models include many realistic details and exclude many others. Their assessment of what was known and unknown regarding factors affecting polio transmission dynamics[29, 30] showed that little was known about waning, other than that there was definite waning after first infections by OPV and WPV in young children. Then using their informal fitting methodology[31] they found that only waning parameters that eliminated all ongoing waning seemed to fit AFPS data in their studies[31]. Consequently, they have used these waning parameters in all

subsequent work. This may not create problems for using their models to inform decisions about control before polio cases have been eliminated. But our analyses show that it is a problem for their studies informing final endgame plans.

The crux of the issue lies in how solid their inference is that waning does not continue beyond a short initial period. We suspect that they found they could fit their model only with parameters that eliminated ongoing waning because they used a highly complex model with many unknown or poorly specified parameters that they fitted using undocumented informal procedures. For example, they adjust contributions of oral and anal transmission to fit patterns from different localities that might have different overall levels of transmission and different contact patterns affecting infection spread. If they set these parameters and infection to paralysis ratios before they fit waning parameters, or before they fit contact pattern parameters, then only waning parameters that do not permit ongoing waning will fit AFPS data. Other similar fitting issues might also explain why they found that only parameters without ongoing waning fit their data. Unfortunately, since they do not make available either their code or the data points to which they informally find fitting parameters, as is needed for good science[89], we could not directly assess their methods to see if our concerns were correct.

4.5.2.2 Institute for Disease Modeling paper

A recent paper[42] proposes a model of waning processes that can be informed by diverse observations including OPV transmission to direct contacts after vaccination, OPV excretion levels and durations after OPV administration given various initial immunity levels, and dose-response observations for OPV and the effects of immunity levels on dose-response. Rather than using direct antibody measures they used an "OPV-equivalent antibody titer" inferred from limited data. They then

fit the inferred antibody level to time using a power law relationship between antibody level and time. While the indirect data used is weak and fits to the power law relationships are correspondingly weak, this work represents the best available data-informed look at waning immunity against polio transmission. The model of waning immunity that emerges from this work is quite different from the model in this chapter. It has continuous waning as in our older paper[86] rather than the one-step waning in this chapter. It encompasses different degrees of boosting of immunity to different levels that depend upon the starting level. Thus, cumulative immunity can be increased significantly by sequential infections. Consequently, the distribution of waned immunity by age will be greater in younger ages and less in older ages than is the case for the waning model in this chapter. The power law relationships inferred between immunity levels and time lead to fast waning early and slower waning later. The decrease in waning over time is considerable. But there seems to be plenty of ongoing waning in their formulation to generate the dynamics we have illustrated that lead to prolonged low-level SC. This waning formulation has the virtue of incorporating different aspects of waning about which it is possible to develop individual level experiments and population level studies that can directly inform model elements. That has two advantages. Fewer parameters must be fit to population levels of infection and more theory can be used when fitting remaining parameters to data. Consequently, parameter identifiability is improved. Finally, waning is conceptualized with greater mechanistic detail. This facilitates the process of inference robustness assessment[75].

4.5.2.3 Polio strains in Pakistan

Alam et al.[4] have described polio strains in Pakistan detected by either ES or AFPS. They find that a wide variety of Type 1 polio strains are still being isolated

and that different strains are geographically limited. They describe long periods where some strains are identified only in ES and not in AFPS. These are in areas with good vaccination. This provides some support to the idea that waning dynamics might be acting to increase potential for prolonged low-level SC. Another aspect of that study, however, may indicate that some aspects of our model do not represent what is happening in Pakistan. ES has found only type 1 and not any type 3 strains. In our model, type 3 strains had more potential for prolonged low-level SC than did type 1. That inconsistency might be explained in different ways. First, it might be that WPV3 has considerably lower transmissibility than type 1. That in turn would diminish the effect of the lower transmissibility of OPV3 compared to OPV1. [29, 30, 31] formulated only slight differences between WPV3 and WPV1. But the data for that formulation are weak. Second, it might be that the silent circulation of WPV3 is very low-level, consistent with our model results. In that case the sensitivity of ES, which is quite low, may not be sufficient to detect prolonged low-level SC. Third, it might be that the dynamics of boosting as formulated in [42] are greater for WPV3 than WPV1. Or the dynamics of waning might be higher.

4.5.2.4 Other studies indicating the potential for prolonged low-level SC

Other studies indicate ongoing waning occurs in the real world. A study of elderly Dutch found that they had durations and levels of OPV excretion after vaccination that were quite long and high[1]. A study in India of excretion after OPV administration showed a loss of immunity between ages 5 and 10[67] that could be explained by ongoing waning of immunity. The potential for SC is illustrated by "orphan" viruses that do not show close genetic links to other recently identified viruses[55, 99]. Viruses might be detected by AFPS or environmental surveillance (ES). Orphan viruses might arise because paralytic polio cases were missed due to in-

adequate AFPS and/or because reinfection transmissions, that are unlikely to cause paralysis, sustained circulation. Both factors might be acting in many situations. The dynamics we illustrate in this chapter show a potential for increasing this latter cause over time in high transmission areas. The most dramatic detection of an orphan virus was the discovery in 2016 of polio cases in Nigeria after more than a year without any detected case in the country. The orphan viruses in this case had not been detected for more than four years[106]. In this case, ES in Northeastern Nigeria detected a cVDPV2 in Nigeria after all OPV2 use had been stopped globally[40]. That stimulated an intensive effort to vaccinate in an area where Boko Haram had severely hampered vaccination and surveillance activities. This effort uncovered the polio cases. ES detected an SC epidemic of WPV1 in Israel in 2013[125]. This SC epidemic occurred because Israel switched to using only IPV in 2005. IPV without prior OPV cannot stop WPV transmission. The resulting SC epidemic was different from the low-level SC we have modeled after long use of OPV. The IPV use caused accumulation of large numbers of children susceptible to infection but not paralysis. Consequently, there were very high levels of infection in Bedouin children in affected villages [84, 112, 113, 114]. Within Israel, administering OPV only to children who had received only IPV successfully eliminated virus transmission without a single paralytic case. The virus came to Israel from Pakistan. It had been previously detected by ES in Egypt but caused no polio cases there. The transmissions carrying virus across these long distances seem more likely to involve older individuals experiencing reinfections. After the Israeli silent epidemic the same virus strain caused polio cases in Syria and Iraq[9, 11]. While the main SC epidemic in Israel does not derive from mechanisms generating SC in the model examined in this chapter, the long path from Pakistan and eventually to Iraq and Syria could be related to those

mechanisms.

4.5.3 Using environmental surveillance (ES) to guide polio eradication end-stage decisions

The dynamics we have elucidated indicate that a strict three year criteria for declaring complete eradication could be inadequate. The origin of the 3 year criterion is largely empirically based on experience in Latin America[24]. Its initial theoretical support used models that assumed acquired immunity never wanes [38, 70]. Recent model analyses assumed that waning only occurs in the first few years after infection[42, 69]. We believe that conditions leading to SC deserve to be studied in a way that allows for a more informed assessment of the chances that SC exists than an arbitrary three years since the last detected polio case.

We propose that ES be analyzed in a new way to assess SC risks. The current state of ES and the plans for expanding it are described in a GPEI report from 2015[47]. ES has already been used to discover long local SC of polio in Pakistan[4]. In Nigeria ES has demonstrated its capacity to guide public health actions[68]. It detected a cVDPV2 that stimulated a special vaccination program after OPV2 use had been ceased in the rest of the world[40]. That in turn led to the discovery of new polio cases[106]. But ES is still not widely used. It covers relatively small populations and therefore would likely miss the low levels of infection generated in our models. Assessing the risks of silent circulation at the time of OPV cessation, therefore, cannot only depend on ES samples. The current use of ES data is just to indicate whether silent circulation exists and not to estimate the chances there is SC.

We propose that analyzing combined ES and AFPS data using transmission models with waning immunity parameters would improve assessment of SC risks. We hy-

pothesize that parameters not identifiable from AFPS or ES data alone can become identifiable when using ES and AFPS data jointly. The reason for this increased identifiability is that AFPS and ES data have different identifiability tradeoffs between parameter values. Given only AFPS data, the effects of parameters that increase waning can be counteracted by increasing both transmission parameters and first infection to paralysis ratios. Given ES data, that is not the case because transmission parameter increases will be directly reflected by ES data. Thus, ES data can pin down parameters that can vary widely when fitting just AFPS data. ES data alone, however, is too sparse to provide evidence against the low levels of polio virus circulation sustained by reinfection seen in our models.

The complementarity of ES and AFPS data should help assess the potential for SC in locations with different conditions such as: history of AFPS polio detections, AFPS adequacy[48], water-sanitation-hygiene (WaSH) conditions[39], and the history of vaccination levels. That will help locate ES where it will be most informative. Currently sites are being located where the most problems in eliminating polio cases have been experienced[68]. But one implication of older age groups sustaining prolonged low-level SC is that populations not exposed to OPV or WPV from kids might have more waned immunity and thus be more likely to sustain reinfection transmission chains given high transmission conditions.

Statistical analyses should be designed to reduce the uncertainty in values of polio model parameters that most affect SC. Inference methods for partially observed stochastic dynamic systems[73] are appropriate for this. They will provide a statistical basis for assessing the probability of SC and thus better inform when OPV should be stopped. The models used should realistically relax the extreme simplifying assumptions of the models used in this chapter. This should be done in an

inference robustness assessment framework[75]. Sequence information could be used to make more precise estimation of the parameters affecting SC by analyzing each strain separately in a single analysis using newly developed methods[74]. This type of statistical analysis becomes increasingly important over time because as the delay time between starting vaccination and eliminating polio cases increases, ever higher vaccination levels can generate ever longer SC if there is not a large sudden increase in vaccination levels. It is not hard to envision a process in which countries with long delays step up their vaccination level to eliminate paralytic cases. As they take more time to do so, the duration of SC increases. If countries can get vaccination coverage and rates above the thresholds that lead to eradication, the global eradication effort will be successful. But because the duration of silent circulation increases with increasing delay time, it will be increasingly difficult to distinguish whether the lack of paralytic cases means eradication is on track or the future may involve a recurrence of polio.

CHAPTER V

Supplementary Material for Dynamics Affecting the Risk of Silent Circulation when Oral Polio Vaccination is Stopped

5.1 Chapter Preface

In Chapter IV (which originally constituted the main text of a paper[77] to which this chapter constituted the supplementary material), I presented a deterministic model of poliovirus transmission that was designed to highlight the potential for all of slow vaccination ramp-up, high transmission potential, ongoing waning, and poor vaccine transmissibility to contribute to a risk of sustained silent circulation. In this chapter, I present a deeper mathematical exploration of that model, including the formulation of an effective reproduction number (R_{eff}) analogue to the Next Generation Matrix basic reproduction number (R_0)[25]. I also present a more detailed justification for some of the analyses performed in Chapter IV, and an in-depth examination of how the rate and depth of waning each contribute to the fraction of transmissions that come from reinfections vs. first infections at the endemic equilibrium. Finally, I compare the simplified model of waning that we used to the more detailed and realistic model presented by Wagner et al.[130].

Symbol	Value	Meaning
n	40	Number of age categories for children under 5
μ	0.02	(Age-independent) Death rate
L	$\frac{1}{\mu}$	Average lifespan
A	[varies]	Average age at (first) infection
λ	$\frac{1}{A} - \mu$	Force of infection operating on fully susceptible individuals, at the pre-vaccination endemic equilibrium
N	10^6	Total size of the modeled population
γ	13/yr	Recovery rate for a first infection with WPV
β	[varies]	Contagiousness for a first infection with WPV
ω	[varies]	Waning rate
θ	[varies]	Ratio of R_0 for OPV to R_0 for WPV
T_{ramp}	[varies]	Duration of vaccination ramp-up
ρ_1	[varies]	Effective rate of OPV vaccination in children under 5 at the end of vaccination ramp-up, before the boost
ρ_2	[varies]	Size of the boost in effective rate of OPV vaccination of children under 5 that occurs at the end of vaccination ramp-up
κ_s	[varies]	Relative susceptibility to (re)infection of individuals with waned immunity, compared to fully susceptibles
κ_d	[varies]	Relative duration (for individuals who recover before they die) of reinfections, compared to first infections
κ_c	[varies]	Relative contagiousness (shedding rate) of reinfections, compared to first infections
κ	[varies]	[see text]

Table 5.1: Parameters

5.2 The effective reproduction number $R_{\mathbf{eff}}$

5.2.1 Tables of parameters and variables

Table 5.1 presents the parameters and variables used in our model that are relevant to the analyses performed in this chapter. For the sake of clarity, and because they are not relevant to these analyses, we omit parameters and variables relating to case detection, silent circulation, and further interventions following the last observed polio case.

In our model, the depth of immune waning following a live virus infection is controlled by a single parameter κ, which denotes the following three values, which we assume for the sake of simplicity to be equal: The relative susceptibility to (re)infection of individuals with waned immunity, compared to fully susceptibles; the relative duration (for individuals who recover before they die) of reinfections,

Symbol	Meaning
S_i	Fully susceptible individuals in age category i
P_i	Partially susceptible (i.e. waned) individuals in age category i
R_i	Fully immune (i.e. unwaned) individuals in age category i
$W_{1,i}$	Individuals in age category i who are experiencing their first live virus infection, with WPV
$W_{R,i}$	Individuals in age category i who are experiencing a second or subsequent live virus infection, with WPV
$O_{1,i}$	Individuals in age category i who are experiencing their first live virus infection, with OPV
$O_{R,i}$	Individuals in age category i who are experiencing a second or subsequent live virus infection, with OPV

Table 5.2: Subpopulations

compared to first infections; and the relative contagiousness (i.e. shedding rate) of reinfections, compared to first infections. Consequently, the effective reproduction number of WPV in a population consisting entirely of individuals with waned immunity would be approximately κ^3 times the basic reproduction number (i.e. the effective reproduction number in a population consisting entirely of fully susceptible individuals). We therefore treat κ^3 as the definition of the waning depth. In Section 5.3, we will consider the distinct role of each of those three values if we do not assume them all to be equal.

Table 5.2 presents the subpopulations represented by compartments in our model.

Table 5.3 presents variables other than subpopulation sizes and time-varying parameters that are present in our model and relevant to the analyses in this chapter.

Symbol	Name in code	Value	Meaning
λ_W	LmbdaW	$\left(\frac{1}{N}\right)\sum_{i=1}^{n+1}(W_{1,i}+\kappa W_{R,i})$	Fraction of the population that is WPV-infected, scaled by fraction of first-WPV-infection contagiousness
λ_O	LmbdaO	$\left(\frac{1}{N}\right)\sum_{i=1}^{n+1}(\theta O_{1,i}+\kappa\theta O_{R,i})$	Fraction of the population that is OPV-infected, scaled by fraction of first-WPV-infection contagiousness
ρ	VaccRtbe	$\begin{cases} 0, & \text{if } t \le 0 \\ \left(\frac{t}{T_{ramp}}\right)\rho_1, & \text{if } 0 \le t \le T_{ramp} \\ \rho_1+\rho_2, & \text{if } T_{ramp} < t \end{cases}$	Vaccination rate for children under five

Table 5.3: Other variables and time-varying parameters

5.2.2 Model equations

The differential equations used in our model are presented below:

$$\frac{d}{dt}\left(S_i\right) = -\left(\beta\left(\lambda_W+\lambda_O\right)+\mu\right)S_i + \begin{cases} \mu N - \left(\frac{n}{5}+\rho\right)S_i, & \text{if } i=1 \\ \left(\frac{n}{5}\right)S_{i-1}, & \text{if } i=n+1 \\ \left(\frac{n}{5}\right)S_{i-1} - \left(\frac{n}{5}+\rho\right)S_i, & \text{otherwise} \end{cases}$$

$$\frac{d}{dt}\left(P_i\right) = -\left(\kappa\beta\left(\lambda_W+\lambda_O\right)+\mu\right)P_i + \omega R_i + \begin{cases} -\left(\frac{n}{5}+\kappa\rho\right)P_i, & \text{if } i=1 \\ \left(\frac{n}{5}\right)P_{i-1}, & \text{if } i=n+1 \\ \left(\frac{n}{5}\right)P_{i-1} - \left(\frac{n}{5}+\kappa\rho\right)P_i, & \text{otherwise} \end{cases}$$

$$\frac{d}{dt}\left(R_i\right) = -\left(\omega+\mu\right)R_i + \gamma W_{1,i} + \left(\frac{\gamma}{\kappa}\right)W_{R,i} + \left(\frac{\gamma}{\theta}\right)O_{1,i} + \left(\frac{\gamma}{\kappa\theta}\right)O_{R,i} + \begin{cases} -\left(\frac{n}{5}\right)R_i, & \text{if } i=1 \\ \left(\frac{n}{5}\right)R_{i-1}, & \text{if } i=n+1 \\ \left(\frac{n}{5}\right)R_{i-1} - \left(\frac{n}{5}\right)R_i, & \text{otherwise} \end{cases}$$

$$\frac{d}{dt}\left(W_{1,i}\right) = -\left(\gamma+\mu\right)W_{1,i} + \beta\lambda_W S_i + \begin{cases} -\left(\frac{n}{5}\right)W_{1,i}, & \text{if } i=1 \\ \left(\frac{n}{5}\right)W_{1,i-1}, & \text{if } i=n+1 \\ \left(\frac{n}{5}\right)W_{1,i-1} - \left(\frac{n}{5}\right)W_{1,i}, & \text{otherwise} \end{cases}$$

$$\frac{d}{dt}\left(W_{R,i}\right) = -\left(\frac{\gamma}{\kappa}+\mu\right)W_{R,i} + \kappa\beta\lambda_W P_i + \begin{cases} -\left(\frac{n}{5}\right)W_{R,i}, & \text{if } i=1 \\ \left(\frac{n}{5}\right)W_{R,i-1}, & \text{if } i=n+1 \\ \left(\frac{n}{5}\right)W_{R,i-1} - \left(\frac{n}{5}\right)W_{R,i}, & \text{otherwise} \end{cases}$$

$$\frac{d}{dt}\left(O_{1,i}\right) = -\left(\frac{\gamma}{\theta}+\mu\right)O_{1,i} + \beta\lambda_O S_i + \begin{cases} \rho S_i - \left(\frac{n}{5}\right)O_{1,i}, & \text{if } i=1 \\ \left(\frac{n}{5}\right)O_{1,i-1}, & \text{if } i=n+1 \\ \rho S_i + \left(\frac{n}{5}\right)O_{1,i-1} - \left(\frac{n}{5}\right)O_{1,i}, & \text{otherwise} \end{cases}$$

$$\frac{d}{dt}\left(O_{R,i}\right) = -\left(\frac{\gamma}{\kappa\theta}+\mu\right)O_{R,i} + \kappa\beta\lambda_O P_i + \begin{cases} -\left(\frac{n}{5}\right)O_{R,i}, & \text{if } i=1 \\ \left(\frac{n}{5}\right)O_{R,i-1}, & \text{if } i=n+1 \\ \left(\frac{n}{5}\right)O_{R,i-1} - \left(\frac{n}{5}\right)O_{R,i}, & \text{otherwise} \end{cases}$$

5.2.3 Approach to defining an effective reproduction number

We define a next-generation matrix effective reproduction number ($R_{\text{eff.}}$) at time t in a fashion analogous to the definition of a next-generation matrix basic reproduction number (R_0) given by Diekmann et al.[25]: $R_{\text{eff.}}$ is the dominant eigenvalue of the next-generation matrix K, whose entries $\{K_{i,j}\}$ represent the expected number of secondary infections into infected subpopulation i that an individual who is infected into infected subpopulation j (at time t) would be expected to produce over the course

of their infection, if the size of all subpopulations were somehow fixed for the duration of that infection. This last constraint is also tacitly present in the definition of an next-generation matrix R_0, being implied by the requirement that the population be at a disease-free equilibrium, apart from an infinitesimal fraction that is infected. Defined in this way, $R_{\text{eff.}} = 1$ if the system is at an *endemic* equilibrium.

5.2.4 Derivation of the equation for the effective reproduction number

In the equations above, none of the dynamics relevant to transmission of WPV depend directly on an individual's age category. Therefore, we can define a smaller set of susceptible and WPV-infected subpopulations by collapsing together subpopulations that only differ with respect to their ages. For understanding the dynamics of this particular sytem, however, it is useful to maintain a distinction between individuals under 5 and individuals who are 5 and older, rather than ignoring *all* distinctions of age:

$$S_1^* = \sum_{i=1}^{n} S_i$$

$$S_2^* = S_{n+1}$$

$$S_3^* = \sum_{i=1}^{n} P_i$$

$$S_4^* = P_{n+1}$$

$$W_1^* = \sum_{i=1}^{n} W_{1,i}$$

$$W_2^* = W_{1,n+1}$$

$$W_3^* = \sum_{i=1}^{n} W_{R,i}$$

$$W_4^* = W_{R,n+1}$$

In this formulation, infections into the the i-th WPV-infected subpopulation (W_i^*) are precisely infections from the i-th susceptible subpopulation (S_i^*). When this is the case, $K_{i,j}$ is simply the product of the expected number of contacts that an individual infected into W_j^* will make with individuals in S_i^* over the course of their infection and the average probability of transmission per contact across those contacts.

Our model has several additional simplifying features as well:

- The contact rate is the same for all individuals.
- Mixing is proportional.
- The contagiousness of an individual infected into W_i^* does not vary over the course of their infection.
- The relative susceptibility of individuals in each susceptible subpopulation S_i^* does not depend on which infected subpopulation W_j^* their potential infector is in.

When all of these are true, the expression for the entries of the next-generation matrix is quite simple:

$$K_{i,j} = \beta_j D_j \sigma_i S_i^*$$

where β_j is the contagiousness of individuals in W_j^*, D_j is the average duration of an infection in W_j^*, and σ_i is the relative susceptibility to infection of individuals in S_i^*. (The values of these parameters are given in Table 5.4, although the following results do not depend on them.)

This implies that all the rows of the matrix K are multiples of each other (specifically, the i-th row is $\frac{\sigma_i S_i^*}{\sigma_j S_j^*}$ times the j-th). Thus, the rank of the matrix K (the maximum number of linearly independent rows of K) is 1. The number of non-zero eigenvalues (counting with multiplicity) of a matrix is equal to its rank; therefore

K has only one non-zero eigenvalue. Consequently, the dominant eigenvalue of K is equal to the sum of the eigenvalues of K. The sum of the eigenvalues of a matrix is its trace (the sum of entries on the main diagonal). Therefore:

$$\begin{aligned} R_{\text{eff.}} &= \sum_{i=1}^{4} K_{i,i} \\ &= \sum_{i=1}^{4} \beta_i D_i \sigma_i S_i^* \end{aligned}$$

$K_{i,i}$ is simply the expected number (under the given assumptions) of secondary transmissions into W_i^* from an individual infected into that same subpopulation. Consequently, we describe it as the contribution to the effective reproduction number of that subpopulation, and obtain the equation given in Chapter IV:

$$R_{\text{eff.}} = R_{\text{eff.First},<5} + R_{\text{eff.Subsequent},<5} + R_{\text{eff.First},\geq 5} + R_{\text{eff.Subsequent},\geq 5}$$

where the various $\{R_{\text{eff.}i}\}$ have the values given in Table 5.5.

Symbol	Value	Meaning
β_j	$\begin{cases} \beta, & j = 1,2 \\ \kappa\beta, & j = 3,4 \end{cases}$	Contagiousness of an individual in W_j^*
D_j	$\begin{cases} \dfrac{1}{\gamma + \mu}, & j = 1,2 \\ \dfrac{\kappa}{\gamma + \kappa\mu}, & j = 3,4 \end{cases}$	Mean duration of infection of an individual in W_j^*
σ_i	$\begin{cases} 1, & i = 1,2 \\ \kappa, & i = 3,4 \end{cases}$	Relative susceptibility of an individual in S_i^*

Table 5.4: Derived variables used to calculate the effective reproduction number

Symbol	Value	Meaning
$R_{\text{eff.First},<5}$	$\left(\frac{\beta}{\gamma+\mu}\right)\sum_{i=1}^{n} S_i$	Contribution of first infections in the under-five age group
$R_{\text{eff.Subsequent},<5}$	$\kappa^3\left(\frac{\beta}{\gamma+\kappa\mu}\right)\sum_{i=1}^{n} P_i$	Contribution of reinfections in the under-five age group
$R_{\text{eff.First},\geq 5}$	$\left(\frac{\beta}{\gamma+\mu}\right) S_{n+1}$	Contribution of first infections in the five-and-older age group
$R_{\text{eff.Subsequent},\geq 5}$	$\kappa^3\left(\frac{\beta}{\gamma+\kappa\mu}\right) P_{n+1}$	Contribution of reinfections in the five-and-older age group

Table 5.5: Contributions to the effective reproduction number

5.3 The effect of waning parameters on the relative contribution of reinfections to the pre-vaccination equilibrium force of infection

5.3.1 Motivation

In Chapter IV, we present four different scenarios for the high-transmission setting: one without waning of immunity following live virus infection, and three with waning of immunity, but varying in the speed and depth of that waning. (We also do this for the low transmission setting, but as discussed in Table 4.1, it was the high-transmission setting that we used to fit our waning parameters for each of the three scenarios with waning; consequently, we focus exclusively on that setting in this section.) All four scenarios have the same average age of first infection (A), and all three scenarios with waning have the same basic reproduction number (R_0) as well.

In this section, we use those two constraints as a starting point in order to understand how the various waning parameters influence the relative contribution of first infections and reinfections to the force of infection at the endemic equilibrium. In this way, we will build a deeper understanding of the interaction between waning depth and waning rate in a broader sense.

Symbol	Meaning
S	Fully susceptible
W_1	Currently first-infected with WPV
W_R	Currently reinfected with WPV
R	Recovered from infection and fully immune
P	Partially susceptible, following waning of immunity

Table 5.6: States of immunity and infection at the pre-vaccination endemic equilibrium

5.3.2 Concepts and symbols

All of the parameters relevant to these analyses are given in Table 5.1. As we noted in Section 5.2, in our model as actually implemented, the depth of immune waning following a live virus infection is controlled by a single parameter κ, which denotes the following three values pertaining to individuals with waned immunity, which we assume for the sake of simplicity to be equal: Relative susceptibility to reinfection (here denoted κ_s), the relative duration of reinfections (κ_d), and the relative contagiousness of reinfections (κ_c) (Table 5.1). However, in order to better show how each of these components of immune waning affect the relative contribution of reinfections to the force of infection at the pre-vaccination endemic equilibrium, we will distinguish them here. At the conclusion of this section, we will show how the results we obtain can be simplified when all three components of immune waning are in fact equal, as is the case for all results presented in Chapter IV.

In this section, we are considering the pre-vaccination equilibrium. Consequently, the vaccine-infected subpopulations all have size 0, and, given our assumptions of homogeneous mixing and age-independent mortality, age is irrelevant. Therefore, the relevant breakdown of the population into subpopulations is that given in Table 5.6

Probabilities and expectations that are derived and used in this section are summarized in Table 5.7

Symbol	Value	Meaning
$\mathbf{P}(X \to Y)$	[see following entries]	Probability for an individual in state X to transition to state Y before dying
$\mathbf{P}(S \to W_1)$	$\frac{\lambda}{\lambda+\mu}$	
$\mathbf{P}(W_1 \to R)$	$\frac{\gamma}{\gamma+\mu}$	
$\mathbf{P}(R \to P)$	$\frac{\omega}{\omega+\mu}$	
$\mathbf{P}(P \to W_R)$	$\frac{\kappa_s\lambda}{\kappa_s\lambda+\mu}$	
$\mathbf{P}(S \to W_R)$	$\frac{\kappa_s\lambda^2\gamma\omega}{(\lambda+\mu)(\gamma+\mu)(\omega+\mu)(\kappa_s\lambda+\mu)}$	
$\mathbf{P}(W_R \to W_R)$	$\frac{\kappa_s\lambda\gamma\omega}{(\gamma+\kappa_d\mu)(\omega+\mu)(\kappa_s\lambda+\mu)}$	Probability for a currently reinfected individual to recover and subsequently be reinfected again, before dying
$\mathbf{E}(W_R)$	$\frac{(\gamma+\kappa_d\mu)(\omega+\mu)(\kappa_s\lambda+\mu)}{(\gamma+\kappa_d\mu)(\omega+\mu)(\kappa_s\lambda+\mu)-\kappa_s\lambda\gamma\omega}$	Expected number of reinfections, conditional on being reinfected at least once
$\mathbf{D}(W_1)$	$\frac{1}{\gamma+\mu}$	Average duration of a first infection
$\mathbf{D}(W_R)$	$\frac{1}{\kappa_d^{-1}\gamma+\mu}$	Average duration of a reinfection

Table 5.7: Probabilities and expectations

5.3.3 Basic Approach

Our model has an exponential population structure, vulnerability to infection that does not depend on age, and homogenous mixing. Consequently, at the endemic equilibrium, the following relationship between the average age of infection, the death rate, and the force of infection holds[26]:

$$A = \frac{1}{\lambda + \mu}$$

Because we fix both the average age of infection and the death rate, this means that the force of infection can be treated as a derived parameter:

$$\lambda = \frac{1}{A} - \mu$$

In order to focus on the effects of the waning parameters, we will begin by treating the (first-infection) contagiousness parameter β as fixed. As we will see in subsequent sections, that parameter will not appear in the equation for the relative contribution of reinfections to the endemic force of infection. This does not mean that it is irrelevant, but rather that it is a free parameter that can be set (as it in fact is, in our model) to the value that produces the correct equilibrium force of infection.

5.3.4 Derivation

5.3.4.1 Equivalence of the relative contribution across the population to the expected relative contribution over a lifetime

Let us consider the expected value of the total contribution that a individual makes to the force of infection over the course of their lifetime, measured in force of infection x time (henceforth, "the total contribution"). We will denote this quantity as Λ. At the endemic equilibrium, the system is stationary. Consequently, at a given point in time, the average contribution of any individual then present in the population

to the force of infection (i.e., $\frac{\lambda}{N}$) is simply the ratio of the total contribution to the average duration of a lifetime:

$$\frac{\lambda}{N} = \frac{\Lambda}{L} = \mu\Lambda$$
$$\therefore \lambda = N\mu\Lambda$$

The endemic force of infection is simply the sum of the endemic force of infection from first infections (λ_{First}) and the endemic force of infection from reinfections ($\lambda_{\text{Subsequent}}$). The same is true of the total contribution and its components (Λ_{First} and $\Lambda_{\text{Subsequent}}$). Consequently, the logic of the paragraph above applies to these components as well:

$$\lambda_{\text{First}} = N\mu\Lambda_{\text{First}}$$
$$\lambda_{\text{Subsequent}} = N\mu\Lambda_{\text{Subsequent}}$$

And thus we obtain the following equation for the relative contribution of reinfections to the endemic force of infection in terms of the total contributions from first infection and from reinfections:

$$\frac{\lambda_{\text{Subsequent}}}{\lambda_{\text{First}}} = \frac{\Lambda_{\text{Subsequent}}}{\Lambda_{\text{First}}}$$

5.3.4.2 The total contribution from first infection

The total contribution from first infection is the product of the probability of being infected before dying, for a fully susceptible individual ($\mathbf{P}(S \to W_1)$); the average duration of a first infection ($\mathbf{D}(W_1)$); and the contagiousness of a first infection (β), divided – because our model is frequency-dependent – by the size of the total

population (N):

$$\begin{aligned}\Lambda_{\text{First}} &= \mathbf{P}\left(S \to W_1\right)\mathbf{D}\left(W_1\right)\beta/N \\ &= \left(\frac{\lambda}{\lambda+\mu}\right)\left(\frac{1}{\gamma+\mu}\right)\beta\left(\frac{1}{N}\right) \\ &= \frac{\lambda\beta}{\left(\lambda+\mu\right)\left(\gamma+\mu\right)N}\end{aligned}$$

5.3.4.3 The total contribution from reinfections

The total contribution from reinfections has a similar form, but is slightly more complicated: It is the product of the probability of being reinfected at least once ($\mathbf{P}\left(S \to W_R\right)$), the average number of reinfections *given* that one is reinfected at least once ($\mathbf{E}\left(W_R\right)$), the average duration of a reinfection ($\mathbf{D}\left(W_R\right)$), and the contagiousness of reinfection ($\kappa_c\beta$), divided (as above) by the size of the total population:

$$\Lambda_{\text{Subsequent}} = \mathbf{P}\left(S \to W_R\right)\mathbf{E}\left(W_R\right)\mathbf{D}\left(W_R\right)\kappa_c\beta/N$$

The probability of being reinfected at least once is the product of probability of being infected at all ($\mathbf{P}\left(S \to W_1\right)$); the probability of recovering before dying, for a first-infected individual ($\mathbf{P}\left(W_1 \to R\right)$); the probability of having one's immunity wane before dying, for a fully-immune individual ($\mathbf{P}\left(R \to P\right)$); and the probability of being reinfected before dying, for an individual with waned immunity ($\mathbf{P}\left(P \to W_R\right)$):

$$\begin{aligned}\mathbf{P}\left(S \to W_R\right) &= \mathbf{P}\left(S \to W_1\right)\mathbf{P}\left(W_1 \to R\right)\mathbf{P}\left(R \to P\right)\mathbf{P}\left(P \to W_R\right) \\ &= \left(\frac{\lambda}{\lambda+\mu}\right)\left(\frac{\gamma}{\gamma+\mu}\right)\left(\frac{\omega}{\omega+\mu}\right)\left(\frac{\kappa_s\lambda}{\kappa_s\lambda+\mu}\right) \\ &= \frac{\kappa_s\lambda^2\gamma\omega}{\left(\lambda+\mu\right)\left(\gamma+\mu\right)\left(\omega+\mu\right)\left(\kappa_s\lambda+\mu\right)}\end{aligned}$$

Similarly, given that an individual is reinfected at least m times, the probability that they are reinfected at least $m+1$ times ($\mathbf{P}\left(W_R \to W_R\right)$) is the product of the probability of recovering before dying, for a reinfected individual ($\mathbf{P}\left(W_R \to R\right)$);

the probability of having one's immunity wane before dying, for a fully-immune individual; and the probability of being reinfected before dying, for an individual with waned immunity:

$$\begin{aligned}\mathbf{P}\left(W_R \to W_R\right) &= \mathbf{P}\left(W_R \to R\right)\mathbf{P}\left(R \to P\right)\mathbf{P}\left(P \to W_R\right) \\ &= \left(\frac{\kappa_d^{-1}\gamma}{\kappa_d^{-1}\gamma + \mu}\right)\left(\frac{\omega}{\omega + \mu}\right)\left(\frac{\kappa_s\lambda}{\kappa_s\lambda + \mu}\right) \\ &= \frac{\kappa_s\lambda\gamma\omega}{\left(\gamma + \kappa_d\mu\right)\left(\omega + \mu\right)\left(\kappa_s\lambda + \mu\right)}\end{aligned}$$

The expected number of reinfections for an individual who is reinfected at least once is therefore:

$$\begin{aligned}\mathbf{E}\left(W_R\right) &= \sum_{i=0}^{\infty}\mathbf{P}\left(W_R \to W_R\right)^i \\ &= \frac{1}{1 - \mathbf{P}\left(W_R \to W_R\right)} \\ &= \frac{1}{1 - \dfrac{\kappa_s\lambda\gamma\omega}{\left(\gamma + \kappa_d\mu\right)\left(\omega + \mu\right)\left(\kappa_s\lambda + \mu\right)}} \\ &= \frac{\left(\gamma + \kappa_d\mu\right)\left(\omega + \mu\right)\left(\kappa_s\lambda + \mu\right)}{\left(\gamma + \kappa_d\mu\right)\left(\omega + \mu\right)\left(\kappa_s\lambda + \mu\right) - \kappa_s\lambda\gamma\omega}\end{aligned}$$

The average duration of a reinfection is:

$$\begin{aligned}\mathbf{D}\left(W_R\right) &= \frac{1}{\kappa_d^{-1}\gamma + \mu} \\ &= \frac{\kappa_d}{\gamma + \kappa_d\mu}\end{aligned}$$

And by substituting all of the above results back into the first equation in this subsection, we obtain the following expression for the endemic force of infection from

reinfections:

$$\begin{aligned}
\Lambda_{\text{Subsequent}} =& \mathbf{P}\left(S \to W_R\right) \mathbf{E}\left(W_R\right) \mathbf{D}\left(W_R\right) \kappa_c \beta / N \\
=& \left(\frac{\kappa_s \lambda^2 \gamma \omega}{\left(\lambda+\mu\right)\left(\gamma+\mu\right)\left(\omega+\mu\right)\left(\kappa_s\lambda+\mu\right)} \right) \times \\
& \left(\frac{\left(\gamma+\kappa_d\mu\right)\left(\omega+\mu\right)\left(\kappa_s\lambda+\mu\right)}{\left(\gamma+\kappa_d\mu\right)\left(\omega+\mu\right)\left(\kappa_s\lambda+\mu\right) - \kappa_s\lambda\gamma\omega} \right) \times \\
& \left(\frac{\kappa_d}{\gamma+\kappa_d\mu} \right) \left(\frac{\kappa_c \beta}{N} \right) \\
=& \left(\frac{\lambda\beta}{\left(\lambda+\mu\right)\left(\gamma+\mu\right)N} \right) \left(\frac{\kappa_s\kappa_c\kappa_d\lambda\gamma\omega}{\left(\gamma+\kappa_d\mu\right)\left(\omega+\mu\right)\left(\kappa_s\lambda+\mu\right) - \kappa_s\lambda\gamma\omega} \right) \\
=& \Lambda_{\text{First}} \left(\frac{\kappa_s\kappa_c\kappa_d\lambda\gamma\omega}{\left(\gamma+\kappa_d\mu\right)\left(\omega+\mu\right)\left(\kappa_s\lambda+\mu\right) - \kappa_s\lambda\gamma\omega} \right)
\end{aligned}$$

5.3.4.4 Conclusion

The relative contribution of reinfections to the endemic force of infection is therefore:

$$\begin{aligned}
\frac{\lambda_{\text{Subsequent}}}{\lambda_{\text{First}}} &= \frac{\Lambda_{\text{Subsequent}}}{\Lambda_{\text{First}}} \\
&= \frac{\kappa_s\kappa_c\kappa_d\lambda\gamma\omega}{\left(\gamma+\kappa_d\mu\right)\left(\omega+\mu\right)\left(\kappa_s\lambda+\mu\right) - \kappa_s\lambda\gamma\omega}
\end{aligned}$$

Thus, the relative contribution of reinfections to the endemic force of infection is directly proportional to (and thus, linear in) the relative contagiousness of reinfections (κ_c). It is also positively dependent on both the relative susceptibility to reinfection (κ_s) and the relative duration of reinfection (κ_d), but is sublinear with respect to each of them (Figure 5.1). However, the degree to which it is sublinear is very different between the two; the dependence on κ_s is strongly sublinear, while the dependence on κ_d is so close to linear that it cannot be visually distinguished from the truly linear dependence on κ_c. This is because the only place that κ_d appears in the denominator is as part of the term $(\gamma + \kappa_d\mu)$, and $\gamma >> \mu$. Finally, the relative contribution of reinfections to the endemic force of infection is also positively, and

sublinearly, dependent on the waning rate (ω) (Figure 5.2).

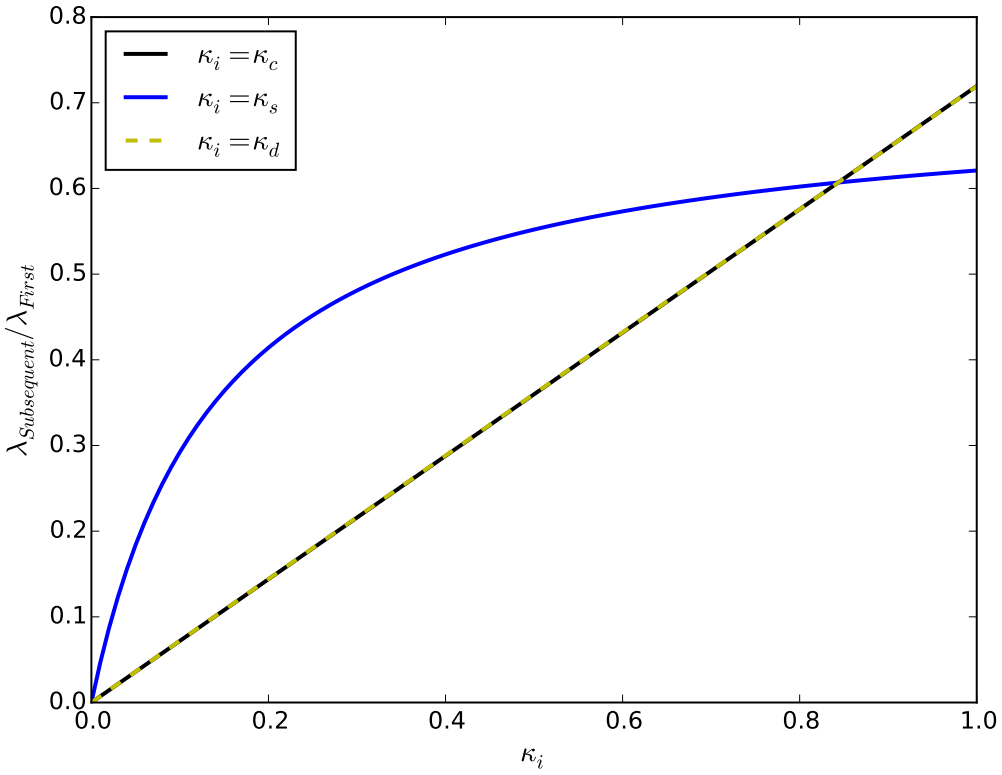


Figure 5.1: Dependence of $\frac{\lambda_{\text{Subsequent}}}{\lambda_{\text{First}}}$ on κ_c, κ_s, and κ_d. In each curve, all parameters other than the κ_i being varied are set to the values they have in the "slow deep" scenario for the high transmission setting in Chapter IV; in particular, both of the other two κ_j are fixed at $0.6^{1/3}$.

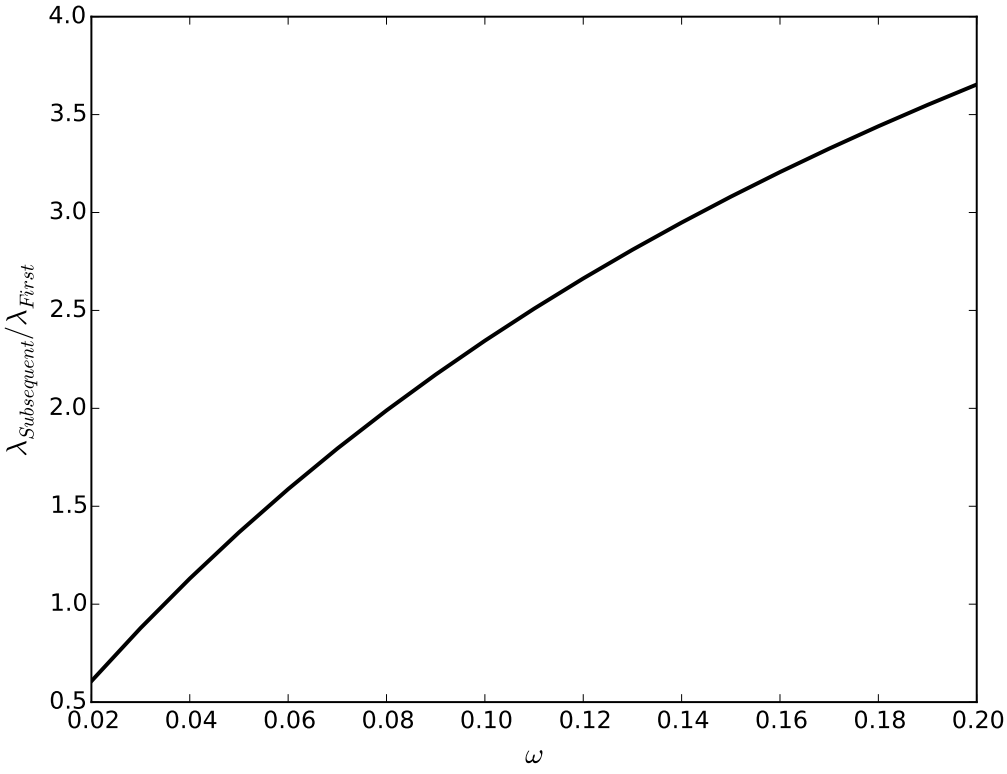


Figure 5.2: Dependence of $\frac{\lambda_{\text{Subsequent}}}{\lambda_{\text{First}}}$ on ω. All parameters other than ω are set to the values they have in the "slow deep" scenario for the high transmission setting in Chapter IV.

5.3.5 Simplification when there is only one parameter κ

In our model as actually implemented, $\kappa_s = \kappa_d = \kappa_c = \kappa$. Consequently, we can simplify the above equation slightly:

$$\frac{\lambda_{\text{Subsequent}}}{\lambda_{\text{First}}} = \frac{\kappa^3 \lambda \gamma \omega}{(\gamma + \kappa\mu)(\omega + \mu)(\kappa\lambda + \mu) - \kappa\lambda\gamma\omega}$$

This expression is clearly superlinear in κ, and just as clearly subcubic. But it is not obvious just from glancing at it where in that range it falls. Numerical tests (Figure 5.3, and others not shown) indicate that it is solidly between quadratic and cubic for the range of parameters considered in any of the scenarios presented in Chapter IV.

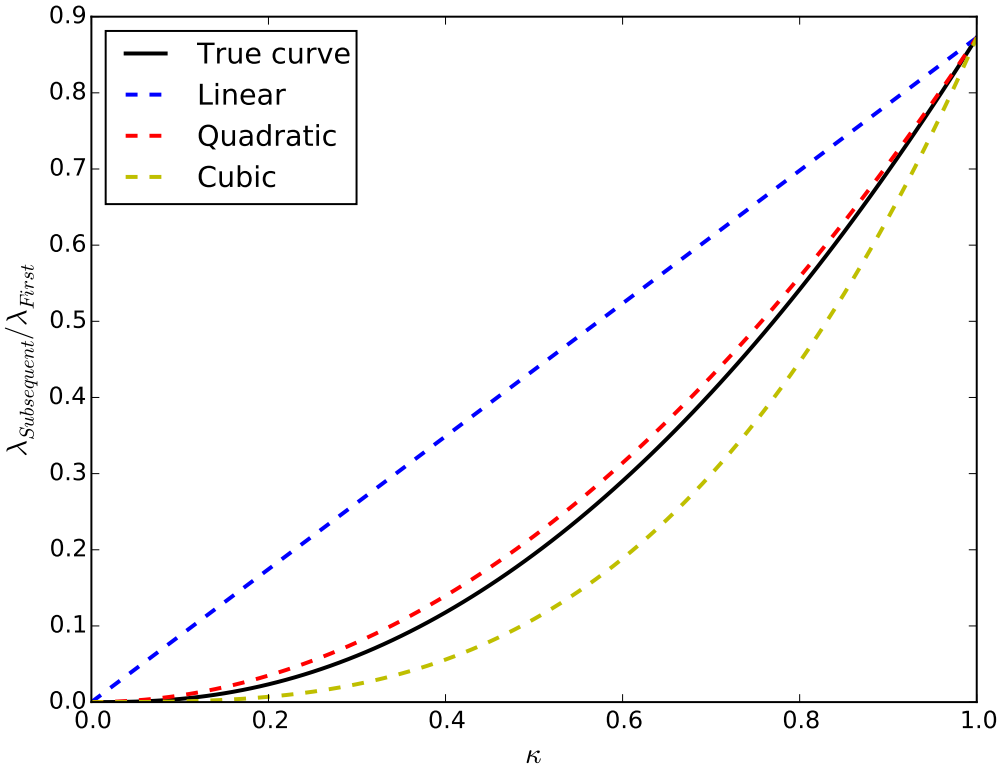


Figure 5.3: Dependence of $\frac{\lambda_{\text{Subsequent}}}{\lambda_{\text{First}}}$ on κ. All parameters other than κ are set to the values they have in the "slow deep" scenario for the high transmission setting in Chapter IV. For comparison, linear, quadratic, and cubic curves having the same two endpoints are shown as dashed lines.

5.4 Waning in the Wagner *et al.* (2014) model

5.4.1 Introduction to the model

One of the more detailed and arguably realistic models of immune boosting and waning following live-virus infection is the one presented by Wagner *et al.*[130], elaborating on the antibody-based model of Behrend *et al.*[12]. Although a full discussion of that model is beyond the scope of this chapter, we will sumarize key aspects of that model, that are relevant to the findings of our model. The relevant parameters, in particular, are summarized in Table 5.8

Symbol	Symbol used in Wagner *et al.* (if different)	Value	Meaning
p_s	P0Inf_s	$\begin{cases} 0.17, & s=1 \\ 0.29, & s=2 \\ 0.13, & s=3 \end{cases}$	Infectivity parameter for strain s
D	PVDose_s	[variable]	Dose of oral polio vaccine (of strain s), in TCID_{50}
A_s	$\text{NAb}_{1,s}$	variable	Effective titer of mucosal antibodies to strain s of poliovirus
M_s	$M_{\text{prime},s}$	$\begin{cases} 5.4, & s=1 \\ 6.17, & s=2 \\ 5.51, & s=3 \end{cases}$	Mean logarithm (base 2) of mucosal antibody titer following a live virus infection with strain s in an immunologically naive individual
σ_s	$\sigma_{\text{prime},s}$	$\begin{cases} 2.2, & s=1 \\ 2.5, & s=2 \\ 2.7, & s=3 \end{cases}$	Standard deviation of the logarithm (base 2) of mucosal antibody titer following a live virus infection with strain s in an immunologically naive individual
τ	τ_{Ab}	0.038	"Neutralization factor for poliovirus by antibodies"[130]
k_d	k_{DI}	0.0469	Parameter governing reduction in average duration of infection due to partial immunity, in $\log_{10}$ (days)/ $\log_2$ (GMT)
k_r	k_{shed}	0.0833	Parameter governing reduction in shedding rate due to partial immunity, in $\log_{10}$ (shed titer)/ $\log_2$ (GMT) (Personal communication, Wagner, 2016)

Table 5.8: Table of parameters in the model of Wagner et. al[130]

5.4.1.1 Probability of infection in the absence of immunity

In the model presented in Wagner *et al.*[130], the probability of infection for an immunologically naive individual challenged with a dose of D TCID_{50} of oral polio vaccine (OPV) of strain s, before accounting for strain interference (which is not a feature of our model) is taken to be:

$$1 - \left(1 + \frac{D}{1 - p_s}\right)^{-p_s}$$

where p_s is a strain-specific infectivity parameter.

In a beta-Poisson dose-response model[44], the susceptibility of targets of infection (which may be individuals, cells, or something in between) varies, with the probability of infection when exposed to a single infectious particle following a Beta(α, β) distribution, and each target of infection is exposed to a (Poisson-distributed) average of $D_{\text{particles}}$ infectious particles. Under some further conditions, the probability of infection under these circumstances can be approximated as:

$$1 - \left(1 + \frac{D_{\text{particles}}}{\beta}\right)^{\alpha}$$

Thus, the dose-response model of Wagner *et al.* is an approximate beta-Poisson model, with individuals as the variably susceptible targets of infection, and with the additional constraint that

$$\alpha + \frac{\beta}{c} = 1,$$

where c is the number of infectious particles per TCID_{50} of OPV. This additional constraint reduces the two-parameter (approximate) beta-Poisson model to a one-parameter model. It is not remarked upon or justified in Wagner *et al.*[130], and is not present in the cited reference[54]. This constraint is clearly carried over from the earlier paper by Behrend *et al.*[12], but is not remarked upon or justified in that paper either.

Presumably, the motivation behind this constraint is that, in the exact beta-Poisson model, $\alpha/(\alpha+\beta)$ is the mean per-particle probability of infection. Thus, by setting $\beta = 1-\alpha$, with $\alpha = p_s$, one appears to obtain the result that the mean per-particle probability of infection is p_s, which would be consistent with the symbol (P0Inf_s) used for that parameter in Wagner *et al.*. But in fact, this would only be true if c were 1 (i.e., if each TCID_{50} of OPV consisted of only a single viral particle), and if that were the case, then the low value of β relative to α would render the approximation to the true beta-Poisson model relatively poor. Nevertheless, for the purposes of this chapter, we will simply take this feature of the model as a given.

5.4.1.2 Acquisition immunity

In the Wagner *et al.* model, immunity to infection is modeled as an unobserved effective titer of mucosal neutralizing antibodies to each of the three serotypes of poliovirus A_s, for $s = 1, 2, 3$. A complete lack of mucosal immunity to serotype s is represented by $A_s = 1$. When an individual is challenged with a dose of D TCID_{50} of OPV of strain s, the probability of infection, taking immunity into account, is:

$$1 - \left(\left(1 + \frac{D}{1-p_s}\right)^{-p_s}\right)^{\frac{1+\tau(A_s-1)\left(1-e^{-1/\tau}\right)}{A_s}}$$

Note that when $A_s = 1$, the above equation indeed reduces to the simpler equation given in the previous section.

5.4.1.3 Intensity of shedding

In the model description presented in the supporting information of Wagner *et al.*[130], the intensity of shedding is implied to be both deterministic (for a given effective mucosal antibody titer) and constant over the duration of infection. However, this contradicts the trajectories of shedding titers over the course of infection

shown in Figure 1(B) in the main text of Wagner *et al.*, which are clearly not constant. The authors have indicated (Wagner, personal communication, 2016) that the presentation in the main text is the correct one, and that the dynamics of shedding are those described in the following paragraphs.

In the absence of immunity, the log_{10} of each individual's peak shedding rate ($S_{i,\text{peak}}$) is selected from a trunctated (at $z = \pm 2$) normal distribution; for each $\log_2(A_s)$, the $\log_{10}$ of $S_{i,\text{peak}}$ is reduced by $k_r = 0.0833$. That individual's shedding rate at a given time since infection ($S_i(t_{inf})$) is then:

$$S_i(t_{inf}) = \begin{cases} 0, & t_{inf} < 1 \\ S_{i,\text{peak}} \left(\dfrac{e^{\frac{-\left(\log\left(t_{inf}\right)-\mu_{LT}\right)^2}{2\sigma_S^2}}}{(t_{inf})\, e^{\frac{\sigma_S^2}{2}-\mu_{LT}}} \right), & 1 \leq t_{inf} < T_i \end{cases}$$

where T_i is the duration of that individual's infection, and

$$\sigma_S = 1.8$$

$$\mu_{LT} = 4.65$$

Although the mean and standard deviation of the distribution of peak shedding titers were not specified, a mean of approximately 4.3 $log_{10}(\text{TCID}_{50}/\text{day})$ can be inferred from Figure 1(B) in Wagner *et al.*. In numerical tests (not shown), our results proved highly insensitive both to the precise value of the mean and to even very large changes (several orders of magnitude) in the standard deviation.

5.4.1.4 Duration of infection

In the absence of immunity, the duration of infection is randomly selected from a normal distribution with a mean of 25 days and an unspecified standard deviation. For each $\log_2(A_s)$, the mean is reduced by a constant k_d, and the standard

deviation is reduced proportionally. In the text of the online Supporting Information of Wagner *et al.*, this reduction is clearly stated to be linear; however, Table S1 of the Supporting Information gives the value of k_d as "0.057 $\log_{10}$(days)/ $\log_2$(GMT)," suggesting that the reduction is proportional. That the reduction is indeed proportional has been confirmed by the authors, who also supplied a corrected value of "0.0469 $\log_{10}$(days)/ $\log_2$(GMT) (Wagner, personal communication, 2016).

5.4.1.5 Priming and boosting of immunity

Priming of immunity in an immunologically naive individual and boosting of immunity in an individual with a previous exposure are handled separately. When an individual with $A_s \leq 1$ experiences a live virus infection with strain s, the logarithm (base 2) of their post-infection antibody titer is randomly selected from a normal distribution with mean M_s and standard deviation σ_s, which vary from strain to strain.

The Wagner *et al.* model also includes boosting of immunity upon live virus infection of an individual with $A_s > 1$. However, for the purpose of simplicity and conceptual clarity, we only examine waning following a single live virus infection in this chapter. Many of the same results apply if waning after multiple infections is considered; however, specifying the precise assumptions about infection and/or exposure history becomes considerably more complex.

5.4.1.6 Waning of immunity

Mucosal immunity and humoral immunity are assume to wane separately. However, the model description in the Supplementary Information indicates that only mucosal immunity is treated as relevant to susceptibility to infection, infection duration, and intensity of shedding, with humoral immunity only being relevant (a) to

protection from paralysis and (b) as a directly measurable proxy (combined with vaccination history) for the unmeasured mucosal immunity. Therefore, in this chapter, we only consider the waning dynamics of mucosal immunity.

Mucosal antibody levels are assumed to undergo simple exponential decay at a rate of either 0.1/yr. (slow waning) or 0.2/yr. (fast waning).

5.4.2 Comparison of waning dynamics to those in our model

For illustrative purposes, the rest of this section focuses on the waning dynamics of the Wagner *et al.* model with parameters for Type 3 poliovirus. All of the major qualitative results also hold for the other two serotypes, although the exact numerical values vary.

5.4.2.1 Acquisition immunity

As seen in Figure 5.4, relative susceptibility to reinfection in our model is comparable to or lower than in the Wagner *et al.* model, in substantial part due to the fact that our model treats individuals who have just recovered from infection as fully immune, while the Wagner model does not. Moreover, the extent of ongoing waning (i.e. continuing waning beyond 5 years after the end of a previous infection) is actually greater in the Wagner *et al.* models than in our models. This suggests that our results regarding the effects of ongoing waning on the potential for sustained silent circulation remain highly relevant when the simplifying assumptions we have made about the nature of waning immunity are relaxed.

5.4.2.2 Intensity of shedding and average duration of infection

Figure 5.5 depicts relative average duration of infection in the various models, while Figure 5.6 covers relative intensity of shedding. These are the same in our model, and are constant over time. This is because, in our model, previously infected

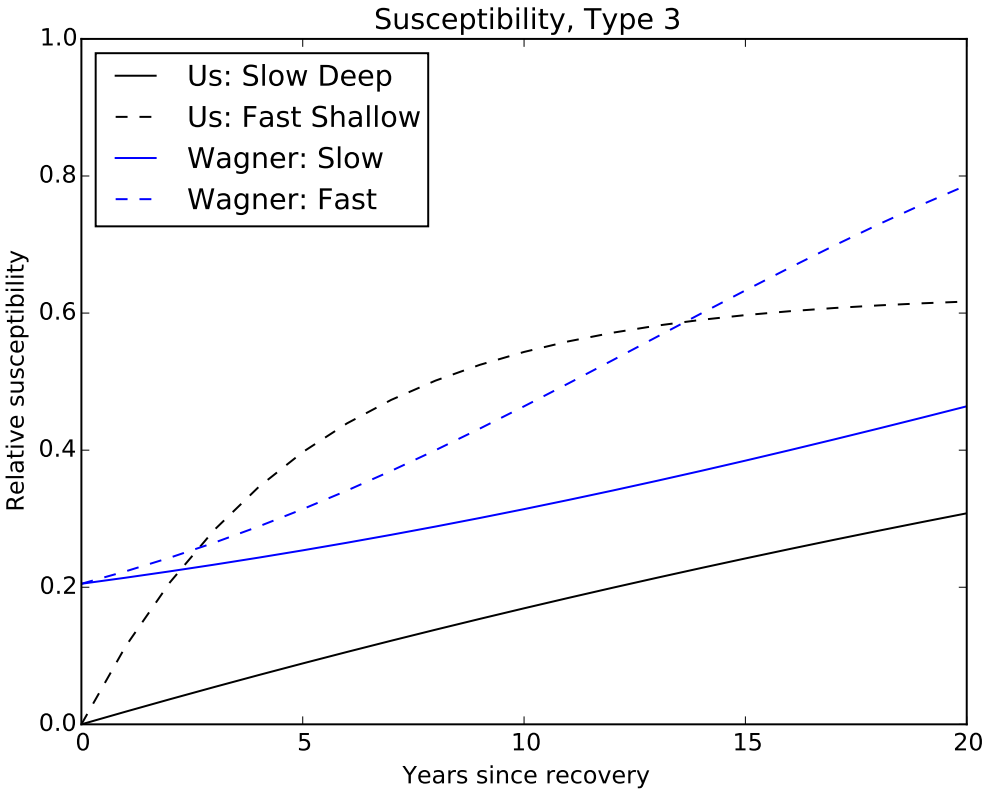


Figure 5.4: Curves showing relative risk of infection with Type 3 poliovirus (compared to an immunologically naive individual) vs. time since previous Type 3 infection in years, under different models of immune waning. Curves for our model are shown in black, and curves for the Wagner *et al.* model in blue. For each model, the solid line represents the sub-model with slowest waning , while the dashed line represents the sub-model with the fastest waning. The curve for our "intermediate" model is not shown. The curves for the Wagner *et al.* model reflect relative risk of infection following a challenge dose of $10^{5.8}$ $TCID_{50}$ of OPV3; the curves for our model indicate relative risk of infection following a cumulative exposure that would produce the same probability of infection in an immunologically naive individual as $10^{5.8}$ $TCID_{50}$ of OPV3 would produce in the Wagner model.

individuals are (at any given time) in one of two homogeneous states: Fully immune, or partially susceptible. Fully immune individuals cannot be infected, and partially susceptible individuals all have the same relative duration of infection and shedding rate. Therefore, the expected duration of infection and shedding rate of reinfected individuals does not depend on the time since first infection; the effects of waning immunity come solely from the increase over time in the fraction of the population that is suceptible.

In contrast, in the Wagner *et al.* model, relative susceptibility, relative duration of infection, and relative shedding rate are all dependent on effective mucosal antibody titer, and so effects of waning immunity are seen in all three ratios. This has a somewhat subtle consequence: In order to appropriately define an average relative duration, it is necessary to weight the distribution of possible antibody titers by the probability of being infected in the first place (given a particular challenge dose). Likewise, in order to appropriately define an average relative shedding rate, it is necessary to weight the distribution of possible antibody titers by the probability of infection times the expected duration. For this reason, the degree to which the curves for relative shedding rate in the Wagner *et al.* model in Figure 5.6 are higher than the corresponding curves for relative duration in Figure 5.6 is greater than would be expected from the difference between k_d and k_r alone. This is because individuals who have lower titers have (1) a higher probability of infection, (2) a longer expected duration of infection if infected, and (3) a higher shedding rate during infection. Therefore, weighting by both the probability of infection and the expected duration of infection (as we do when calculating the average relative shedding rate) would result in a higher average than weighting by only the probability of infection (as we do when calculating the average duration of infection), even if the probability

distributions were identical apart from this weighting.

As regards both duration of infection and shedding rate, the potential for transmission from reinfections is substantially higher in the Wagner *et al.* model than in ours, although it should be stressed that this observation is made in the context of immunity generated by only a single previous infection, and will naturally be less true the greater the number of previous infections (including vaccinations that succeed in producing a live virus infection). More robust, however, is the observation that ongoing waning is again stronger in the Wagner *et al.* model than in ours.

5.4.2.3 Overall

In Figure 5.7, we compare the various models with respect to the overall relative potential of previously infected individuals to contribute to transmission, as measured by relative probability of infection times relative average duration of infection times relative shedding rate, analogous to Figure 4.2, which provides that comparison with respect to our models only. Again, we see that both the relative potential for reinfections to contribute to transmission, measured in this way, and the extent to which waning is ongoing are greater in the Wagner *et al.* model than in ours. This strongly suggests that our conclusion that ongoing waning can greatly increase the potential for extended silent circulation is robust to relaxation of our simplifying assumptions about the dynamics of waning of mucosal immunity following live virus infection.

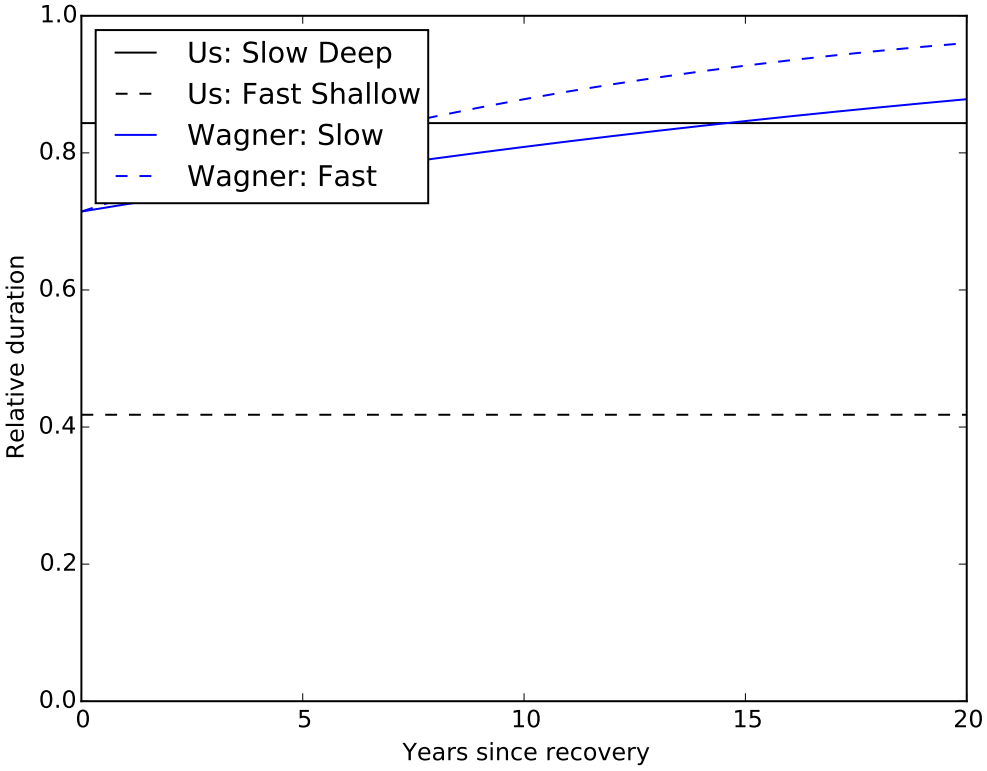


Figure 5.5: Curves showing relative average duration of infection while infected with Type 3 poliovirus (compared to an immunologically naive individual) vs. time since previous Type 3 infection in years, under different models of immune waning. Curves for our model are shown in black, and curves for the Wagner *et al.* model in blue. For each model, the solid line represents the sub-model with slowest waning , while the dashed line represents the sub-model with the fastest waning. The curve for our "intermediate" model is not shown. The curves for the Wagner *et al.* model reflect relative risk of infection following a challenge dose of $10^{5.8}$ $TCID_{50}$ of OPV3; the curves for our model are the same regardless of challenge dose, for reasons discussed in the text.

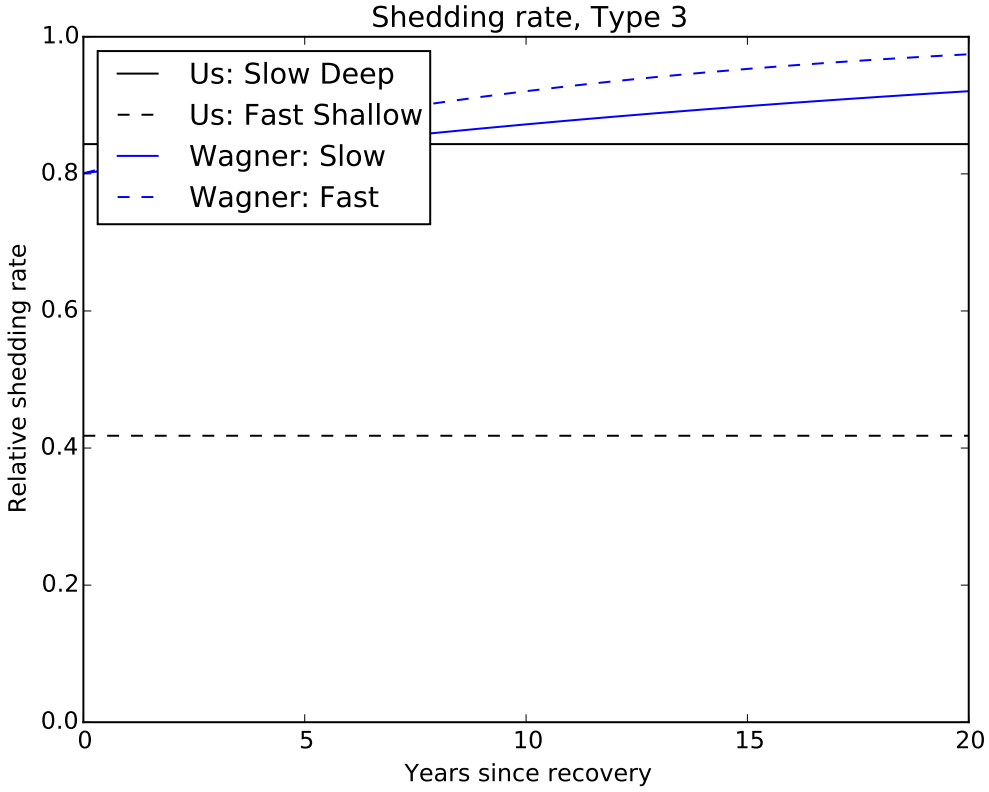


Figure 5.6: Curves showing relative intensity of virus shedding while infected with Type 3 poliovirus (compared to an immunologically naive individual) vs. time since previous Type 3 infection in years, under different models of immune waning. Curves for our model are shown in black, and curves for the Wagner *et al.* model in blue. For each model, the solid line represents the sub-model with slowest waning , while the dashed line represents the sub-model with the fastest waning. The curve for our "intermediate" model is not shown. The curves for the Wagner *et al.* model reflect relative risk of infection following a challenge dose of $10^{5.8}$ $TCID_{50}$ of OPV3; the curves for our model are the same regardless of challenge dose, for reasons discussed in the text.

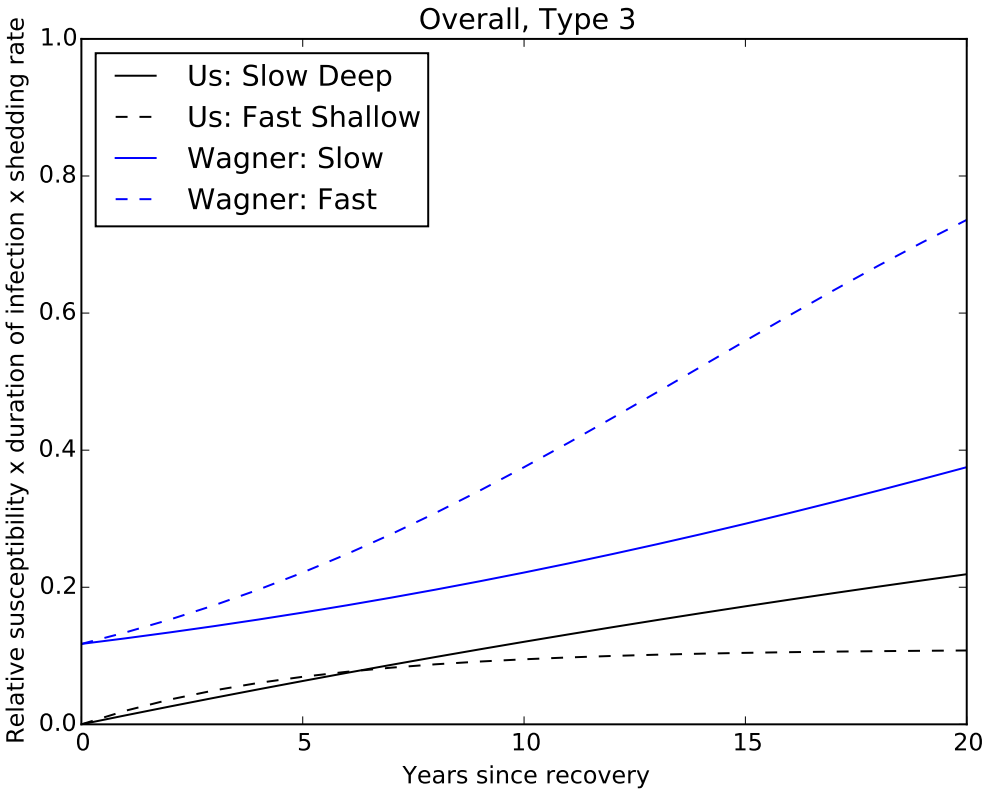


Figure 5.7: Curves showing relative probability of infection with Type 3 poliovirus times relative average duration of infection times relative shedding rate (compared to an immunologically naive individual) vs. time since previous Type 3 infection in years, under different models of immune waning. For our model, this is equivalent to the "population extent of waning" shown in Figure 4.2. Curves for our model are shown in black, and curves for the Wagner *et al.* model in blue. For each model, the solid line represents the sub-model with slowest waning , while the dashed line represents the sub-model with the fastest waning. The curve for our "intermediate" model is not shown. The curves for the Wagner *et al.* model reflect relative risk of infection following a challenge dose of $10^{5.8}$ $TCID_{50}$ of OPV3; the curves for our model are the same regardless of challenge dose, for reasons discussed in the text.

CHAPTER VI

Hybrid Deterministic-Stochastic Model for Understanding the Effects of Ongoing Waning of Immunity on the Risk of Sustained Silent Poliovirus Circulation

6.1 Introduction

In Chapters IV and V, we constructed and analyzed a model of poliovirus transmission dynamics designed to improve our understanding of how waning dynamics affect the risk of prolonged silent circulation, potentially ending in a rebound epidemic. This model provided several valuable insights. But it has limitations as well. Two major ones both relate to its deterministic compartmental nature, in somewhat different ways.

One is that, being based on a large-sample continuous approximation to discrete population dynamics, it does not naturally handle eradication well, as this involves the infected population reaching zero. With continuous population and first-order death and recovery, this will never actually happen. In our deterministic model, this was handled *ad hoc* by declaring "stochastic elimination" once the expected number of infected individuals drops below 1. This is a reasonable way of dealing with the problem, but it is not ideal.

The other is more broadly that silent circulation as a whole is characterized by random processes: Its beginning is marked by the moment at which the last para-

lytic case occurs before an extended period without paralytic cases, which depends not only on when transmissions of wild poliovirus to fully susceptible individuals ("wild first infections") occur, but also on which wild first infectees actually become paralyzed, and even on *when* in the course of their infections paralysis occurs; its end is marked by either the moment at which the last infected individual recovers (or, in a system with environmental transmission, the somewhat subsequent moment at which the last already-shed wild poliovirus ceases to be viable) or the moment at which the next paralytic case occurs. None of these are actually deterministic. Moreover, they are all rare on the time scale that we care about, and so a deterministic model provides a poor approximation to this stochastic reality. And just as our model does not "naturally" generate a time of eradication, so too, it does not naturally generate a period without cases, nor a moment at which cases return; infinitesimal fractions of cases are occuring continually at all times at which WPV is present in the population.

In Chapters IV and V, we solved this problem much as we did the other: We declared silent circulation to begin at the moment at which the expected number of (deterministic) cases in the past year first dipped below one, and to end in case return (if elimination had not then already occured) at the next subsequent moment when it climbed back above one. Again, this was workable, but not ideal even from the perspective of obtaining a single result.

But the broader problem is that a single result is not ideal for this sort of question anyway. Realistically, even if we knew the exact nature of transmission dynamics, and the exact current conditions (excluding any quantum mechanical "hidden variables" that might exist), there would be a distribution of possible outcomes. Even if some central measures are known, the degree of uncertainty implied by the rest of the

distribution still matters. If the median duration of silent circulation is low, it remains important whether or not the standard deviation is high. If the median outcome of silent circulation is elimination, it matters whether elimination occurs with 99% probability, of 51%. And so forth.

In this chapter, I address these two problems, by incorporating some stochastic elements into the model.

6.2 Methods

For this chapter, we extended the deterministic compartmental model from Chapters IV and V (previously published together as a paper in the journal Epidemics[77]) to incorporate stochastic components based on techniques used in a 2015 paper by Michael Famulare[41]. Although the full details of the models used in those papers are given there, we summarize them below. The parameters used in each of these models, and in the hybrid model that we construct from a synthesis of the two, are given in Table 6.1.

6.2.1 The deterministic model developed in the preceding chapters

In Chapters IV and V, we used a deterministic compartmental model of moderate complexity. While not as elaborate as some models[31, 42, 130], that model nevertheless features a number of distinct phenomena. Reinfection is present in the model, along with waning of immunity following a previous infection. This waning is abstracted as a single transition from a fully immune state to a partially susceptible state, in which an individual is once again capable of being infected, but with a reduced probability. If infected, a partially susceptible individual also recovers more quickly than a fully susceptible one, and sheds less virus per gram of stool over the course of infection.

Symbol	Value from Chapter IV	Value in Famulare	Meaning
N	10^6	[Not used]	Total size of the modeled population
μ	0.02/yr	[Not directly set]	(Age-independent) Death rate
γ	13/yr	[Not directly set]	Recovery rate for a first infection with WPV
γ^*	$\gamma + \mu$	$\sim$ Unif($\frac{1}{30\text{days}}, \frac{1}{7\text{days}}$)	Total rate of removal from the infected population
β	$\frac{R_0}{\gamma+\mu}$	[Not used directly]	Contagiousness for a first infection with WPV
κ	[varies]	[Not used, implicitly 0]	Cube root of waning depth
$\lambda_{1,1}(t)$	$\frac{\beta S(t)}{N}$	$\sim$ N($\gamma^*, 3.44$/yr)	Transmission rate to fully susceptibles, per first-infected
$\lambda_{R,1}(t)$	$\frac{\kappa\beta P(t)}{N}$	[implicitly 0]	Transmission rate to partially susceptibles, per first-infected
$\lambda_{1,R}(t)$	$\kappa\lambda_{1,1}(t)$	[inapplicable]	Transmission rate to fully susceptibles, per reinfected
$\lambda_{R,R}(t)$	$\kappa\lambda_{2,1}(t)$	[inapplicable]	Transmission rate to partially susceptibles, per reinfected

Table 6.1: Parameters and selected variables relevant to this chapter from the deterministic compartmental model of Chapters IV and V, and from the stochastic model in a 2015 paper by Michael Famulare[41]. The $\{\lambda_{i,j}\}$ do not appear directly in Chapters IV and V, where a more typical force-of-infection formulation is used, but this formulation is fully equivalent, and more suitable for the purpose of this chapter's model.

The model also distinguishes between OPV and WPV, but only tracks dynamics for one serotype at a time, and treats succesful vaccination and secondary infection with OPV identically. Consequently, there are 7 basic states that an individual can be in: Fully susceptible; infected for the first time, with WPV; infected for the first time, with OPV; partially susceptible; reinfected, with WPV; reinfected, with OPV; and fully immune.

In addition to these immune dynamics, the model also tracks age, dividing the population into 41 age groups: Children under the age of 5 are separated into a series of 1.5-month age groups, while individuals over the age of 5 are all lumped together. Mortality is assumed, for the sake of simplicity, to be constant with respect to age, at a rate of 0.02/year. The only effect of age is that children under the age of 5 are subject to (direct) vaccination, while older individuals are not. Thus, age is a source

of comparatively high computational complexity, relative to conceptual complexity.

As a deterministic compartmental model, this model is suitable for efficient simulation of dynamics when the number of (actual) individuals whom the abstraction assigns to each compartment is sufficiently large. However, it may fail to account for stochastic effects — not only when the total population is small, but also when key subpopulations are small, such as when prevalence is very low.

The beginning of "silent circulation" (silent, that is, with respect to AFP surveillance) is declared when the expected number of cases (based on the number of new first infections and the case-to-first-infection ratio) in the previous year falls to 1, in a total population of 1 million. If the total *prevalence* of *all* infections falls further, below 1 per million, "stochastic elimination" is declared as the outcome of the simulation; if, instead, the expected number of cases since the begining of silent circulation reaches 1, then "case return" is declared.

6.2.2 The stochastic model of Famulare

The paper by Famulare[41] presents a very different model, one in which the effects of partially sucessful polio eradication campaigns are assumed to primarily take the form of rendering some regions (of whatever size) incapable of sustaining polio transmission, while others are affected not at all. In a sense, this is an opposite assumption to that made by the model in Chapter IV, which uses a simplifying assumption of mass action within a homogeneous-mixing population. In both cases, different parameter values are provided for each serotype, but dynamics are only simulated for one serotype at a time.

Using this assumption, the Famulare paper[41] presents a stochastic model that only tracks infected individuals. Only a single class of infecteds (implictly: first infections only) is included. There are only two types of transition in this model:

an increase in the number of infecteds due to transmission, and a decrease in the number of infecteds due to recovery:

$$a_{W_1 \to W_1+1} = \lambda_{1,1}(t) W_1$$

$$a_{W_1 \to W_1-1} = (\gamma^*) W_1$$

where γ^* is the "recovery rate" in the sense of being the rate of leaving the infected state by any means; because the model only tracks infected individuals, it does not matter whether this occurs by true recovery or by death. Thus, in terms of how we defined parameters in Chapter V, $\gamma^* = \gamma + \mu$, as indicated in Table 6.1.

Simulation is performed using a standard (direct) Gillespie algorithm[45]. Cases of paralytic poliomyelitis occur (randomly, with probability equal to the estimated case-to-first-infection ratio) on recovery; this is numerically an extremely good approximation to cases occuring at a random time during the course of infection.

$\gamma*$, the recovery (plus death) rate, is drawn once for each simulation run from a uniform distribution on the interval $[\frac{1}{30\text{days}}, \frac{1}{7\text{days}}]$, and is kept constant throughout that run. Given γ^*, an initial value for W_1 is calculated to match the expected time between the second-to-last and last observed cases with the actual length of that interval for the serotype in question, based on published case-count data from Nigeria.

The expected growth rate of the infected population at a given point in time is allowed to vary by setting $\lambda_{1,1}(t) = \gamma^* + N(0, \sigma)$ where σ is fitted from case count data from Nigeria, and is equal to 3.44yr^{-1}. (In practice, λ is constrained to be ≥ 0, for obvious reasons.) This distributon is resampled at three-month intervals.

This continues until either $W_1 \to 0$ or a case is detected.

6.2.3 Synthesis and model description

Each of these approaches has unique strengths that the other does not. The model used in Chapter IV, unlike Famulare's[41], includes waning dynamics and transmissions from individuals with waned immunity, which can not only increase the ratio of infections to paralysis, but can also cause that ratio to vary over time. Famulare's model, unlike the one used in Chapter IV, simulates small-population dynamics in ways that account for the importance of stochastic fluctuations, and its probabilistic nature allows for obtaining a range of outcomes with an associated probability distribution. Compared to stochastic approaches that simulate the entire population, its approach of only directly simulating infected individuals greatly reduces the number of events (and even more so, the number of *classes* of events) that must be simulated, increasing computational efficiency. We sought to formulate a hybrid model that would combine the strengths of these two models.

In order to do this, while retaining much of the computational efficiency benefits of a deterministic model, we made use of the following observation: In the fully deterministic model of Chapters IV and V, during and around the period of silent circulation, the ratio of susceptibles of a particular class (fully susceptible or partially susceptible (i.e., waned)) to currently-infected individuals who were infected from that class consistently exceeds 84:1 for all parameter sets; for some parameter sets, it consistently exceeds *20,000*:1.

This has two consequences. First, explicit stochastic simulation of infecteds is computationally cheap compared to explicit stochastic simulation of the entire population. Second, stochastic fluctuations in the number of individuals infected in a given time period (during that interval) can have a very large proportional effect on the number of infecteds, while having only a very small effect on the number

of susceptibles. Consequently, explicit stochastic simulation only of infecteds, while treating the number of individuals in each susceptible or fully immune category as deterministic, is a reasonable approximation to stochastic simulation of the entire system.

Therefore, for a given parameter set, we begin by running a fully deterministic simulation of the system as described in Chapters IV and V. The code used to do this has been converted from Berkeley Madonna to R, and some of the equations have been restructured for improved clarity, but no changes have been made to the underlying dynamics of the system. We simulate this trajectory out to a point 50 years after the beginning of vaccination. We then record the size of all compartments in 0.01-year intervals, from the first point at which the total size of the infected population is 300 or less (out of a total population of 1 million), until the end of the simulation. The starting point of the recording, which is also used as the starting point for our stochastic simulations, consistently occurs between the end of vaccination ramp-up and the declaration of the beginning of silent circulation in the purely deterministic model.

Because our stochastic simulation only directly deals with currently infected individuals, and we assume (as in Chapters IV and V) that vaccination of a currently-infected individual has no effect, the stochastic portion of our model is not (directly) affected by vaccination. Because vaccination is the only age-dependent effect in our model – mortality being assumed to be constant with respect to age for the sake of simplicity – this means that we can collapse subcompartments across age categories. As noted above, this eliminates a quite substantial source of computational complexity in the fully deterministic model.

Similarly, we can ignore the vaccine-infected and fully immune compartments,

as it is not possible (in our model) to be infected with WPV directly out of them. Therefore, the compartments whose deterministic values are relevant to stochastic dynamics during potential silent circulation of WPV are thus reduced to four: Fully susceptibles, first-infecteds, partially susceptibles, and re-infecteds. Moreover, for the infected compartments, only the deterministic values at the start of the 10-year interval are relevant, for initializing the model; after that, the values will be allowed to vary stochastically.

The stochastic simulation therefore proceeds as follows: First, the numbers of first-infecteds (W_1) and reinfecteds (W_R) are initialized to their deterministic values at the starting point of the similation (i.e., the first point in the deterministic simulation where their sum has been reduced to 300). Then, at each time point of .01 years (a granularity selected in order to ensure that changes in $\{\lambda_{i,j}\}$ between time points are too small for further reductions to meaningfully affect the results), propensities are calculated according to the following formulae:

$$
\begin{aligned}
a_{(W_1,W_R)\to(W_1+1,W_R)} &= \lambda_{1,1}(t)W_1 + \lambda_{1,R}(t)W_R \\
a_{(W_1,W_R)\to(W_1-1,W_R)} &= (\mu+\gamma)W_1 \\
a_{(W_1,W_R)\to(W_1,W_R+1)} &= \lambda_{R,1}(t)W_1 + \lambda_{R,R}(t)W_R \\
a_{(W_1,W_R)\to(W_1,W_R-1)} &= (\mu+\frac{\gamma}{\kappa})W_R
\end{aligned}
$$

where W_1 and W_R have their current stochastic values, and $\{\lambda_{i,j}(t)\}$ are calculated from the current numbers of fully susceptible and partially susceptible individuals, taken from the deterministic simulation (Table 6.1).

The time-to-next-event and nature of the next event are then selected in accordance with a standard direct Gillespie procedure. If the selected time is less than or

equal to the next observed time point, then W_1 or W_R is updated (and a possible case triggered in the case of infection into W_1, as in Chapter IV), as is t, and the simulation proceeds. If it is greater than the next observed time point, the simulation instead advances to the next observed time point, recalculates propensities in accordance with the new values for S and P, and draws a new time-to-next-event and nature of that event accordingly. This is a permissible approach due to the memoryless property of an exponential distribution[45].

A silent circulation interval is defined as either (a) the interval between two consecutive polio AFP cases or (b) the interval between the last polio AFP case and the end of all local circulation (i.e., true elimination). Clearly, if there is an interval of type (b), it must be the *last* silent circulation interval, given that our model does not include reintroduction from outside the simulated population.

Given this fact, we can classify all silent circulation intervals into three categories: If the interval ends in elimination, it is an elimination interval. If the interval ends in a case, but elimination occurs within the 20-year period following the start of the simulation, then it is a blip interval. If neither of the above is true, then it is a true return interval (i.e., it is either immediately or eventually followed by a rebound epidemic). Looking at the 20-year period as a whole, we are principally interested in two things: The longest duration of a silent circulation inteval, and the nature of that interval – with one caveat: If the longest interval is an elimination interval, but there is a blip interval that is over 3 years long, we focus on the latter over the former. The rationale here is that we wish to focus on the extent to which silent circulation has the potential to create major problems for global polio eradication efforts.In order to answer that question correctly, it is important to take into account the existing GPEI threshold of 3 years without a detected polio case to declare elimination of endemic

circulation in a given country[87]. Thus, case return that occurs after an interval of three or more years is case return that has the potential to occur after the country (or multi-country epidemiological block, such as Afghanistan and Pakistan) is declared polio-free. The potential effects of this are severe enough to warrant considering this a worse threat to polio eradication than even a substantially longer interval that results in eradication.This is essentially a high-severity local threat. Below the three-year threshold, local severity is expected to be much lower, because elimination will not yet have been certified when silent circulation ends. Therefore, the long-term potential for export to other, currently polio-free, countries from a very long interval — regardless of how that interval terminates locally — takes precedence. For ease of reference, we will henceforth refer to the interval selected in this manner as the "worst interval." Once the worst interval has been selected, we treat blip and true return intervals identically.

6.3 Results

Using the model described above, we obtained results for all 3 serotypes of poliovirus, for scenarios of no waning, fast-shallow waning, and slow-deep waning, for apparent R_0s of 10 and 15, and for final effective vaccination rates from 0 to 1.5 per year, as defined in Chapter IV. From these, we made plots that contained curves for (1) the probability that the length of the worst interval is less than or equal to a given duration, (2) the probability that the worst interval is of a given type (case return or elimination) and that its length is less than or equal to a given duration, and (3) the fraction of worst intervals whose length is less than or equal to a given duration that end in case return, rather than in elimination. A representative example, for Type 3 poliovirus with slow-deep waning, an apparent R_0 of 15, and a final effective

vaccination rate of 1.00 yr^{-1}, is shown as Figure 6.1.

Examining the dotted red line, indicating the fraction of resolved intervals that terminated in case return, reveals an interesting phenomenon: Early resolutions are almost all case return, because the probability of a single very early paralytic case is greater than the probability of the numerous very early recoveries that are required in order to achieve elimination in a very short time frame. As the duration of silent circulation intervals increases, those intervals have an increased probability of ending with elimination. But past about 3 years or so, the fraction of intervals ending in case return begins to increase again, as waning immunity, in the absence of large-scale WPV transmission, increases the susceptibility of the population.

In Figure 6.2, we present a set of similar plots for various strain and waning scenario combinations, retaining an apparent R_0 of 15 and a final effective vaccination rate of 1.00 yr^{-1}; hence, the lower right panel of Figure 6.2 is simply a scaled-down version of Figure 6.1. Comparing the different panels of Figure 6.2, we can make several observations. Going down each column, we see that, as in Chapter IV, both introducing waning and shifting from fast-shallow waning to slow-deep waning increase the potential for prolonged silent circulation. Going across each row, we see that, also in agreement with Chapter IV, Type 3 exhibits the greatest potential for sustained silent circulation, followed by Type 1, with Type 2 exhibiting the least, despite having the highest first-infection-to-paralysis ratio (IPR). This pattern matches the extent to which each wild strain is more transmissible than the corresponding vaccine strain; hence it agrees with the result in Chapter IV that vaccine transmissibility has a greater impact on the potential for sustained silent circulation than IPR.

There is also a clear interaction between these two effects, such that the poten-

tial for prolonged silent circulation in the Type 3, slow-deeping waning scenario is substantially higher than one would expect by comparing the no-waning scenario for Type 1 or Type 2, the slow-deep waning scenario for that same serotype, and the no waning scenario for Type 3. This interaction is also reflected in the fact that an increased probability of very long silent circulation intervals ending in case return, relative to those of intermediate length, can be seen in the Type 3, slow-deep panel, but not in either the Type 1, slow-deep panel, nor in the Type 3, fast-shallow panel. Finally, we can also see that slow, deep waning not only increases the potential for prolonged silent circulation, but also the probability for prolonged silent circulation to end in case return.

Simply comparing outcomes across a single effective vaccination rate is not necessarily the most informative approach, however. As can be seen in Figure 4.4, the effective vaccination rate that results in the longest potential silent circulation varies widely between scenarios. Therefore, in order to better understand the range of possible outcomes, we took these results and made summary plots showing quantiles (specifically, the medians of quintiles, i.e. the 10th, 30th, 50th, 70th, 90th percentiles) of the time-to-event distribution vs. total effective vaccination rate. To begin with, we present a single such plot, for Type 1 poliovirus and slow-deep waning, in Figure 6.3. This plot also indicates, by means of the colors used, the fraction of intervals in each quintile that are case-return rather than elimination, the same information that is indicated by the dotted red lines in Figures 6.1 and 6.2.

Several features are notable. Unsurprisingly, higher total effective vaccination rates result in lower fraction of case return. Careful examination of the locations on each of the curves of the transitions between colors, reflecting shifts in the fraction of intervals in each quintile ending in case return, reveals the same phenomenon noted

above, in relation to the dashed red lines in Figures 6.1 and 6.2: Very short silent circulation intervals almost all end in case return, then as the duration increases, the probability of ending in elimination increases. Finally, sufficiently late outcomes may shift back toward case return due to waning of WPV-induced immunity. This latter effect can be seen by comparing the rate of case return at the peak of the highest-quintile curve to the rates of case return for the middle three quintiles at the same effective vaccination rate.

To compare results across strains and waning scenarios, we construct a paneled figure (Figure 6.4) corresponding to Figure 6.3 in the same way that Figure 6.2 corresponds to Figure 6.1. Comparing the various parameter sweeps contained in Figure 6.4 confirms the observation made in relation to Figure 6.2, that both increased levels of ongoing waning and decreased OPV transmissibility result in both increased potential for sustained silent circulation, and an increased probability that such circulation, if it occurs, will end in case return.

The fact that the longest durations of silent circulation are clearly associated with Type 3 differs from the pattern seen in Figure 4.9, where the very longest durations occur in the narrow peaks for Type 2. This is, however, consistent with the narrowness of those peaks, and the much greater breadth of the long-duration peaks for Type 3: The introduction of stochastic effects into a deterministic system has a tendency to smooth out sharp transitions, which can result in very narrow peaks disappearing altogether.

Finally, two contrasting points: The median outcome, indicated by the solid curve in each plot, is acceptable (in the sense of being less than 3 years in duration) for almost all parameter sets — only for Type 3 and slow-deep waning is this not the case for any effective vaccination rates, and even there, most of the range of effective

vaccination rates examined results in a median duration of less than three years. This is consistent with empirical observations, as one would expect: The median empirical observation should lie near the middle of a probability distribution. But our empirical observations in very high-transmission settings come from a relatively small number of countries, which makes it difficult to use them to make inferences about the full probability distributions. And there is substantial probability mass at high — and in some cases, *extremely* high — durations of silent circulation for a wide range of parameter values for which the median outcome is good. For example, for Type 3, slow-deep waning, and an effective vaccination rate of 1.0/yr., the median duration of silent circulation is around 2.5 years. But there is a full 10% probability of silent circulation lasting over twice that long, almost twice the GPEI standard of 3 years for certifying elimination! This means that, while the most likely outcome is very likely to be that elimination or case return (as the case may be) will occur prior to certification of national elimination, there is nevertheless a grave risk that it will not, in one or more countries. But this risk varies substantially between waning scenarios (as well as between serotypes). This increases the importance of developing a fuller understanding of waning dynamics, as well as of improving and extending environmental surveillance.

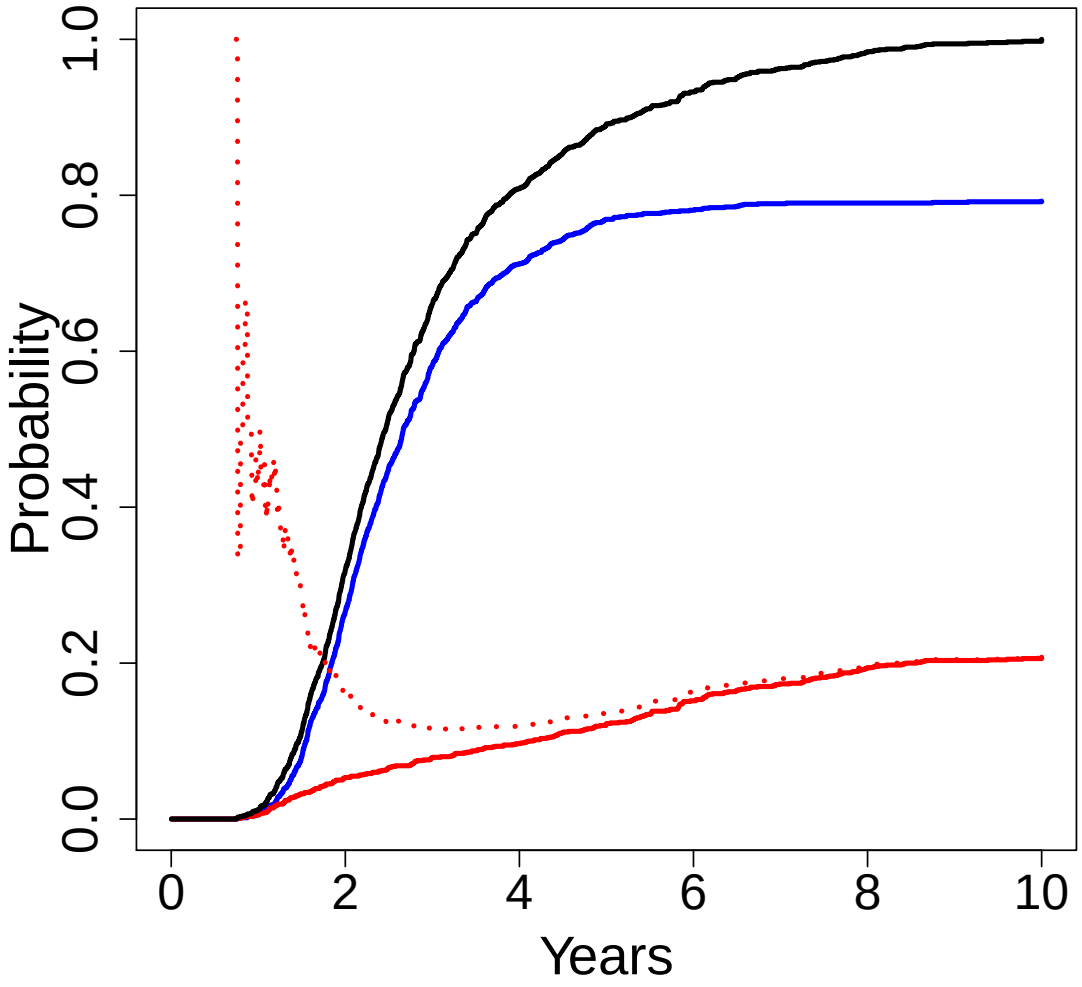
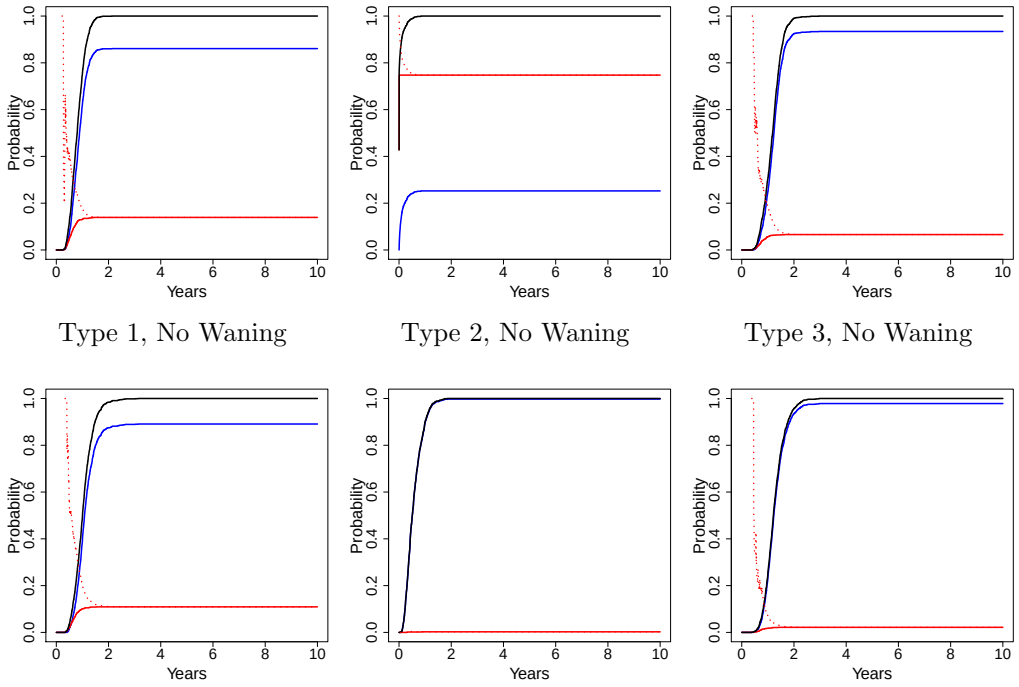


Figure 6.1: Distribution of outcomes for Type 3 poliovirus, with slow deep waning, an apparent R_0 of 15, and an effective vaccination rate of 1.00/year. The black solid curve represents the cumulative distribution of worst interval durations, regardless of the nature of those intervals. Separating those intervals into those that end in case return and those that end in elimination results in the red and blue solid curves, respectively. Thus, for any given x-value, the y-values of the blue and red curves sum to the y-value of the black curve. The red dotted curve indicates the fraction of worst intervals of duration less than or equal to a given value that end in elimination, i.e., it is the ratio of the y-value of the red solid curve to the y-value of the black solid curve.



Type 1, No Waning | Type 2, No Waning | Type 3, No Waning

Type 1, Fast-shallow Waning | Type 2, Fast-shallow Waning | Type 3, Fast-shallow Waning

Type 1, Slow-deep Waning | Type 2, Slow-deep Waning | Type 3, Slow-deep Waning

Figure 6.2: Distribution of outcomes as in Figure 6.1, by poliovirus type (columns), and immune waning dynamics (rows). All results are presented for an apparent R_0 of 15 and an effective vaccination rate of 1.00/year. The manner of results within each panel is the same as in Figure 6.1 (which is identical to the bottom right panel): The black solid curve represents the cumulative distribution of worst interval durations, regardless of the nature of those intervals. Separating those into those that end in case return and those that end in elimination results in the red and blue solid curves, respectively. Thus, for any given x-value, the y-values of the blue and red curves sum to the y-value of the black curve. The red dotted curve indicates the fraction of worst intervals of duration less than or equal to a given value that end in elimination, i.e., it is the ratio of the y-value of the red solid curve to the y-value of the black solid curve.

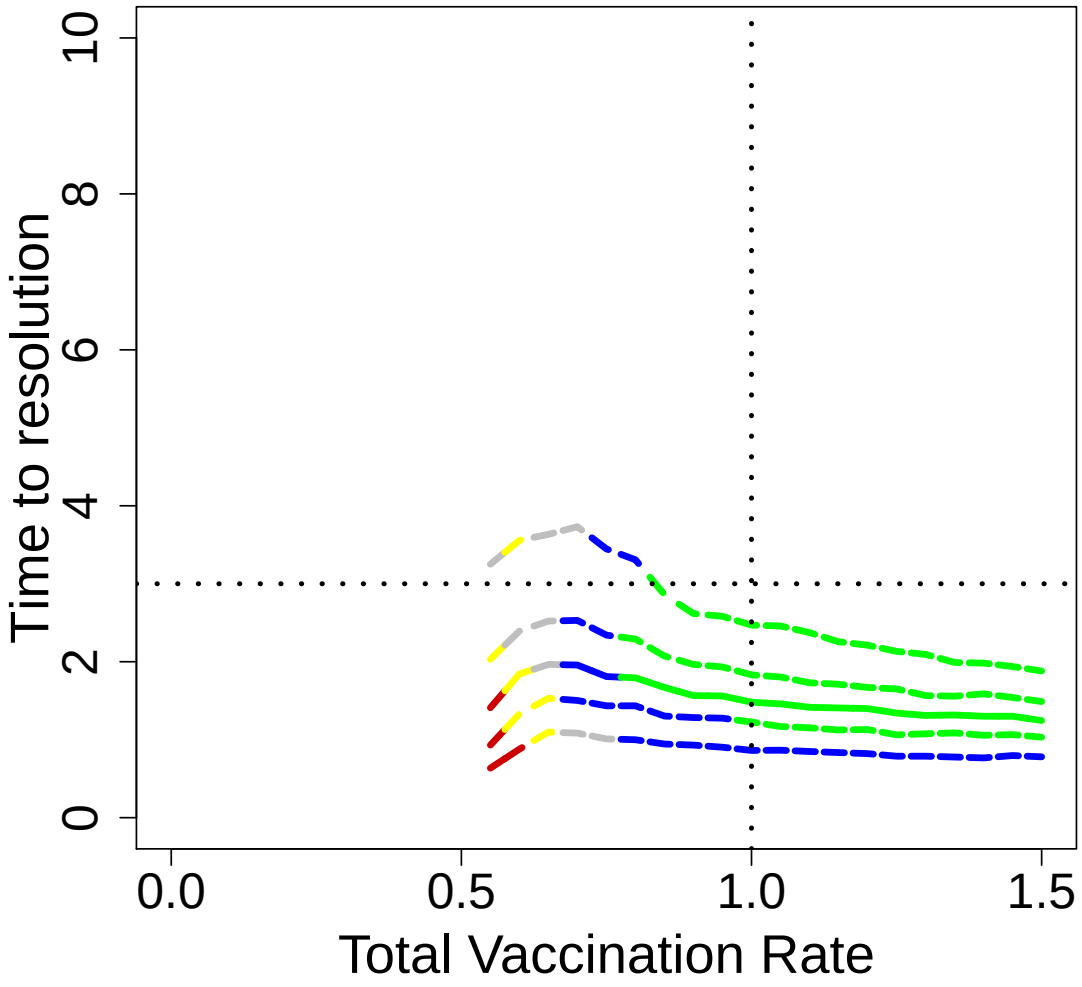
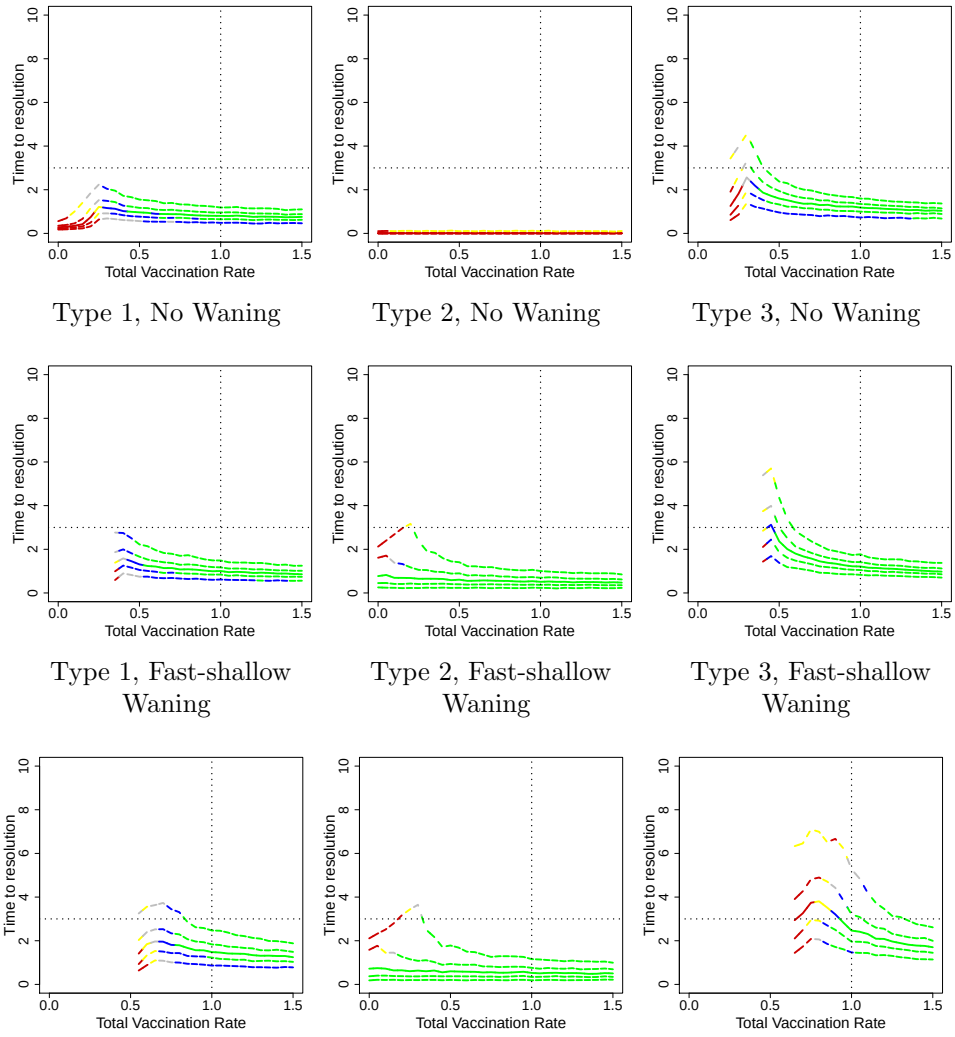


Figure 6.3: Summary of quintiles of outcomes by effective vaccination rate. Each curve represents one quintile of worst interval durations; for a given effective vaccination rate (x-value), the y-value of that curve represents the median duration of intervals in that quintile (e.g., for the second curve from the bottom, the y-value indicates the 30th percentile of interval durations for that effective vaccination rate). Thus, a single vertical "slice" amounts to a summary of the cumulative distribution curve of ("worst") interval durations for a single parameter set. In particular, the slice at x = 1.00 yr^{-1}, denoted by the vertical dotted black line, contains a summary of the information contained in the solid black curve of the Type 1, slow-deep panel of Figure 6.2. The color of each curve at a given point reflects the fraction of intervals in that quintile that end in case return: Red indicates 80-100% case return, yellow indicates 60-80%, grey indicates 40-60%, blue indicates 20-40%, and green indicates 0-20%. For emphasis, the third-quintile curve (i.e., the one whose y-values represent median durations) is solid, while the other four are dashed. The horizontal dotted black line in each panel indicates a duration of 3 years, equal to the GPEI standard for certifying eradication[87].



Type 1, No Waning | Type 2, No Waning | Type 3, No Waning

Type 1, Fast-shallow Waning | Type 2, Fast-shallow Waning | Type 3, Fast-shallow Waning

Type 1, Slow-deep Waning | Type 2, Slow-deep Waning | Type 3, Slow-deep Waning

Figure 6.4: Distribution of outcomes as in Figure 6.3, by poliovirus type (columns), and immune waning dynamics (rows). The presentation of results within each panel is the same as in Figure 6.3 (which is identical to the bottom left panel): Each curve represents one quintile of worst interval durations; for a given effective vaccination rate (x-value), the y-value of that curve represents the median duration of intervals in that quintile (e.g., for the second curve from the bottom, the y-value indicates the 30th percentile of interval durations for that effective vaccination rate). Thus, a single vertical "slice" amounts to a summary of the cumulative distribution curve of ("worst") interval durations for a single parameter set. In particular, the slice at x = 1.00 yr^{-1}, denoted by the vertical dotted black line, contains a summary of the information contained in the solid black curve of the Type 1, slow-deep panel of Figure 6.2. The color of each curve at a given point reflects the fraction of intervals in that quintile that end in case return: Red indicates 80-100% case return, yellow indicates 60-80%, grey indicates 40-60%, blue indicates 20-40%, and green indicates 0-20%. For emphasis, the third-quintile curve (i.e., the one whose y-values represent median durations) is solid, while the other four are dashed. The horizontal dotted black line in each panel indicates a duration of 3 years, equal to the GPEI standard for certifying elimination[87].

6.4 Conclusion

In Chapters IV and V, we presented an analysis of several aspects of both poliovirus transmission dynamics and poliovirus control efforts that affect the risk sustained silent circulation during the polio eradication endgame. The model used in those chapters had sufficient complexity to reflect key areas of uncertainty, while also making use of simplifications designed to make it easier to understand the causal mechanisms underlying patterns in our results. One of those simplifications was the use of a deterministic compartmental model. In this chapter, we relaxed some of the assumptions underlying that simplification by incorporating partial stochasticity into that model. This allowed us to retain a large portion of the computational efficiency of the original model, while more accurately handling phenomena like stochastic die-off and discrete, individual cases of paralytic poliomyelitis. Moreover, it allowed us to replace the single, deterministic outcome that the earlier model generated for each parameter set with a range of more or less likely possible outcomes, thereby acknowledging the inherent role of chance in transmission dynamics when the total number of infected individuals is low.

On those questions which the purely deterministic model was capable of addressing, our basic conclusions remained much the same: Ongoing waning and low vaccine transmissibility are both important causal factors that increase both the risk of sustained silent circulation and the risk of such circulation ending in case return. The role of vaccine transmissibility, in particular, is reflected in the fact that our model predicts the greatest risk of sustained silent circulation for Type 3, and the lowest for Type 2, despite Type 2 having the highest ratio of first infections to polio AFP cases. There is also evidence of a positive interaction between these two effects.

An area in which the hybrid model greatly outshines the purely deterministic one is in understanding the multiple basic types of uncertainty which are relevant when attempting to formulate appropriate strategies for minimizing the risk that sustained silent circulation will result in rebound and/or reintroduction epidemics following certification of eradication by the GPEI. One type of uncertainty is uncertainty in parameter values. In the case of waning parameters, that uncertainty itself has already been discussed in Chapter IV, and its implications regarding the potential for sustained silent circulation can be readily perceived using either model.

But there is an additional form of uncertainty which purely deterministic models are structurally incapable of capturing, and that is uncertainty as to what outcome will in fact occur, even given a complete and perfect knowledge of all the parameters characterizing a transmission system. In Chapter IV, we noted that both the intensity of transmission in a particular setting prior to the start of vaccination, and the speed with which vaccination is ramped up to sufficient levels to drastically reduce the frequency of paralytic infection have a strong effect on the potential for sustained silent circulation. In that chapter, we suggested that this means that empirical results observed during polio eradication efforts in Latin America are not sufficient to dismiss the possibility of silent circulation being sustained beyond three years in the higher-transmission settings of the remaining endemic countries. This conclusion continues to hold. But the further implication of the additional form of uncertainty captured in the results found in this chapter is that even empirical results observed during eradication efforts in high-transmission settings may not be sufficient to confidently predict the results that will be observed in other high-transmission settings, unless they are supported by analyses that are capable of making key inferences about the role of reinfection in sustaining transmission. For many plausible param-

eter sets, both sustained silent circulation and the absence thereof have substantial probability.

Moreover, these two forms of uncertainty naturally interact with each other: For any given parameter set, there will be range of outcomes corresponding to some particular confidence level, such as 95 or 99%. But the breadth of that range (and not merely its location) depends heavily on what those parameters are, as can be seen in Figures 6.3 and 6.4. Thus, some aspects of what we do not know render us unable to confidently say how important other aspects of what we do not know really are. Consequently, continuing to expand environmental surveillance is necessary both to in order to better understand polio transmission dynamics, and to directly safeguard eradication efforts.

In this chapter, we have shown the existence of substantial regions of plausible parameter space where the median outcome was silent circulation lasting 3 years or less, but the probability of substantially longer silent circulation, often ending in case return, remained unacceptably high. The implication of this is that even our empirical experience from eradication efforts in very high-transmission settings, such as Northern India, is not sufficient to provide the necessary level of confidence in our predictions of outcomes in the remaining endemic countries upon which to base potentially high-risk policy decisions.

As with all models, the model used in this chapter has a number of limitations. Broadly, these fall into two categories. The first is those that stem from the fact that, while we have relaxed several common simplifying assumptions in the modeling of poliovirus transmission dynamics in this and the preceeding two chapters, we have still made numerous simplifying assumptions of our own. The two most prominent of these are homogeneous mixing and a simplified and somewhat unrealistic population

age structure, implied by our use of a constant death rate for all age groups. One-step waning is another. A lack of environmental transmission is yet another, although this is somewhat offset by the assumption of homogeneous transmission. The other broad category, and a rather more pervasive one, is the relative lack of hard data used in formulating this model. Because of this, we are, fundamentally, pointing out realistic uncertainties, not resolving them. We are pointing out important and often overlooked questions, but are limited in our ability, using solely the results presented here, to provide answers.

The next steps suggested by this work fall into two basic categories: Modeling and field work. As implied by the procedural framework of inference robustness assessment, these must not be treated as separate and unrelated sets of tasks to be completed, but must be allow to inform each other in a continuous process[75]. For modeling, one broad area of work that remains to be done is realistically relaxing the simplifying assumptions that are built into this model. There are an endless array of ways in which this can be done, and more real-world data is needed in order to identify those that are most likely to be practically useful. But one that is especially likely to be of value is adding more realistic contact patterns than the homogeneous mixing used here.

Another is adding environmental surveillance (ES) to the model. Adding ES to acute flacid paralysis surveillance (AFPS) has already proven highly valuable for detecting circulation of poliovirus that would not have been detected through AFPS alone, most dramatically in the "silent" Israeli reintroduction outbreak of 2013–2014[8, 113]. This has been accomplished in an essentially empirical fashion. But modeling work, constrained by what data already exists, could be used to determine how to do this more effectively. Because political considerations may constrain the

ways in which ES can realistically be deployed, it is important that this work be informed by ongoing communication with policymakers[75]. In addition, we conjecture that combining ES and AFPS in a suitably informed fashion will allow us to distinguish between parameter sets that differ greatly in their potential for sustained silent circulation but that cannot be distinguished by AFPS alone.

Another, more purely theoretical, next step is to develop a better understanding for why the locations of the peaks in median duration of silent circulation in the hybrid model (specifically, the Type 3 panels of Figure 6.4) are noticeably shifted, relative to the location of the corresponding peaks (namely, the 20 year delay peaks in panels E, F, and H of Figure 4.4) in the purely deterministic model. It is unsurprising that the peaks themselves would be lower. But it is not clear why they should occur at lower effective vaccination rates. Possibly the answer will prove to be irrelevant to the real-world inferences that we seek to make. But possibly not. And even if the reason proves to be unimportant in the context of polio eradication efforts, it may still provide insights into non-obvious ways in the choice of modeling approach influences results that may be useful in future research.

One possible first step toward developing such an understanding is to break down further the shift from a deterministic to a (partially) stochastic handling of infected individuals. Comparing the model presented in this chapter to that presented in Chapter IV, all of the following have been changed from deterministic to stochastic: (1) Infection, (2) recovery from infection, and (3) which first infections result in paralytic cases, both (a) for the purpose of determining when silent circulation begins, and (b) for the purpose of determining when silent circulation ends in case return. All of these can in principle be separated, although for some of them, mathematical subtleties will be involved.

Another possible first step, in some sense opposite to the one above, is to simulate fully stochastic trajectories for the system, including all individuals, susceptible, immune, WPV-infected, and OPV-infected. If these trajectories generate results that look more like the fully deterministic results than the hybrid results, then that suggests that the shift seen in the hybrid results is a misleading modeling artifact, a product of some sort of subtle bias. This seems unlikely, but it is not impossible. And while understanding modeling artifacts generally does little to create an improved understanding of actual transmission systems, it can sometimes to much to create an improved understanding of how to avoid creating similar artifacts in future work.

As for field work, there are two major next steps to be taken. We have already noted the importance of continuing to expand environmental surveillance. In order to obtain the full benefits of doing so for understanding the underlying transmission dynamics, and even to simply obtain the greatest possible benefits for detecting silent circulation in the moment, strategies for doing so must be informed by careful analysis. But even if analysis should be unavoidably delayed, expansion on an empirical basis remains better than no expansion at all, in terms both of increasing our ability to detect silent circulation directly, and of generating data that can be analyzed in the future. Such data is unlikely to be ideal for that purpose. But will still constitute a starting point, and a vast improvement over not having such a starting point.

Secondly, work must continue on the creation of improved polio vaccines[80]. This is not directly relevant to the inference issues described herein. But it has the potential to perhaps make some of them less pressing.

CHAPTER VII

Conclusion

Mathematical modeling of transmission dynamics can never represent *all* features of those dynamics, nor should it in general aim to do so. Instead, its goal must be to represent all those features which are salient for the purpose of making particular inferences that are relevant to public health decisions. Which features are salient is not intrinsic to the system itself, considered in a vacuum, but is specific to the particular inferences related to that system that we as modelers are seeking to make. The way to identify these features is through an inference robustness assessment[75], a continuous cycle in which we realistically relax simplifying assumptions, determine whether our inferences are robust to that relaxation, and then, based on the answer to that question, either gather additional data or relax additional assumptions.

As inference robustness assessment is in principle an infinite cycle, performing a complete robustness assessment for a particular inference is outside the scope of this dissertation. Instead, I have examined in various fashions the consequences of realistically relaxing particular simplifying assumptions as regards the transmission dynamics of two very different pathogens, HIV and poliovirus. I have shown that key inferences about these systems are not robust to such relaxation, and have highlighted some ways in which additional data might be gathered in order to constrain the

plausible parameter space.

Despite the many differences between HIV and poliovirus, they have certain key features in common. In both cases, it is difficult or impossible to detect most infections near the time when they occur. Partly, this is a natural consequence of the natural history of the diseases that each causes, although the aspects of those natural histories that generate this consequence are different for each of the two.

HIV produces a relatively brief acute infection that frequently produces no overt symptoms at all, and that rarely requires medical treatment even when symptoms occur. Moreover, even if infected individuals do get tested during this phase of infection, HIV antibody tests (the most common, although not only, type of HIV test) can often return a negative result[111]. Following acute infection, asymptomatic chronic infection can last for years. Thus, although almost all HIV infections in the United States and other developed, Western nations will *eventually* be detected, most of them will not be detected until long after they first occurred.

In contrast, most poliovirus infections will never be detected at all. The vast majority will produce only gastrointestinal symptoms, and a large majority of those that produce additional symptoms will only produce non-specific ones, such as fever and muscle aches. And all, excepting only those that occur in individuals with rare immunodeficiencies, will be fully defeated by the host's immune system within a relatively short period of time. Consequently, only the tiny fraction that cause paralytic symptoms are likely to be detected (as discrete individual infections, as opposed to merely the presence of infection within a broader population that environmental surveillance can detect) in any fashion other than by pure chance.

Beyond their respective natural difficulties in being detected, HIV and polio share an additional obstacle: What individual-level data does exist, despite the difficulties

in collecting it, is often quite difficult to obtain from the organizations (primarily government agencies) that collect it, for essentially political reasons. Again, the exact causes differ: In the case of HIV, agencies are generally concerned about the potential for such data to be used in lawsuits by infected individuals against their infectors, whether due to concerns that this may disincentivize testing, general aversion to controvery and negative public attention, or both. In the case of polio, the causes are somewhat less obvious, but seem to be primarily a fear of embarassment of one's country on the world stage at a national level, and a mixture of a desire to maintain good relations with national governments and simple bureaucratic turf-defending at an international level.

Despite differences in the details, this obstacle presents similar problems for inference robustness assessment for both polio and HIV. Most obviously, it increases the difficulty in performing such assessments at all, aggravating the problems created by the *natural* difficulties in obtaining adequate data created by the biology of the viruses themselves. Somewhat more subtly, it causes the principal or even exclusive curators of key transmission data to be individuals who generally have little expertise in modeling. Because an effective inference robustness assessment requires close collaboration between modelers, policymakers, and data gatherers[75], this sort of siloing is a major obstacle to answering not only questions of interest to modelers, but questions of great importance to policymakers as well.

Sometimes, data-sharing agreements are crafted that allow access to data for small numbers of specific modeling groups. Such agreements are, strictly speaking, better than nothing. But they still cripple the process of improving our collective ability to make robust inferences about key policy decisions, relative to open sharing of deidentified data with everyone, and of data that by its nature cannot be deidentified, with

any group that is capable of meeting the necessary ethical and security standards.

Some simplifying assumptions are encoded into models as fixed values for explicit parameters. But many more are encoded through certain parameters not being explicitly included in the model at all, but tacitly assumed, through their omission, to have some particular value (most frequently 0 or 1). In order for the latter set of assumptions to be realistically relaxed, they must first be recognized. What is "obvious" upon a cursory examination to one modeler is often difficult or even impossible for another to realize on their own, even after careful examination. This is true even when both are of comparable experience and skill, due to differences in background, or simply in the patterns of thought that come naturally to each. Consequently, limiting those who can engage with the data to a handful of specific groups ensures that only a fraction of possible insights into simplifying assumptions to which inferences may not be robust will occur. In some cases, this limited set of insights will be sufficient. But in others, it undoubtedly will not. As it is impossible, in general, to predict in advance when this will be the case, this provides a strong argument for making as much data as possible available as broadly as possible, subject only to such ethical constraints as may exist and be insurmountable.

In Chapter II, I presented and analyzed a model highlighting the role of episodic risk in HIV transmission systems at a fixed homogeneous transmission potential and fixed fraction of that potential corresponding to transmissions during early infection. This model included multiple risk groups, assortative mixing (conceptualized in terms of multiple mixing sites), and a simplified version of the natural history of HIV infection consisting of a short, highly-infectious acute phase and a longer, less-infectious chronic phase. This analysis demonstrated a strong and non-monotonic effect of episodic risk on both the fraction of transmission occuring during early in-

fection and the endemic prevalence. Among the implications of these results was that the expected effect of interventions promoting less risky behavior is not robust to the possibility of those interventions increasing the extent to which risk heterogeneity is episodic by promoting an initial reduction in individuals' risky behaviors, followed by a later rebound.

In Chapter III, I extended this model to incorporate (a simplified version of) universal test and treat (UT&T), and tweaked the parameterization used to characterize episodic risk in order to both better reflect key mathematical features of the system and more readily generalize to more realistic scenarios having more than two risk groups. Using this model, I presented an analysis of the effects of episodic risk and heightened transmissibility during early infection when the *endemic prevalence* is fixed, rather than the homogeneous transmission potential. This more realistically reflects what we do and do not know about the state of the HIV epidemic among American men who have sex with men. In this context, I found that the fraction of transmissions that occurred during early infection depended non-monotonically on the degree to which risk heterogeneity was episodic, but that this effect was in practice dominated, when it came to the effect on the effective treatment rate required to achieve control of transmission, by the strong and monotonic decrease in the basic reproduction number with increasingly episodic risk. The contrast between this result and some of the results obtained in Chapter II shows how crucial a proper framing of the inferences to be made is. This reemphasizes the importance of close collaboration between modelers and policymakers, with their respective areas of expertise, and of getting as many eyeballs on a problem as possible.

In Chapter IV, I presented and analyzed a model highlighting the role of ongoing waning in the potential for poliovirus transmission systems to generate sustained

silent circulation at a fixed average age of first infection and fixed prevalence of first infections at the end of an initial vaccination ramp-up. This model included differentiation of the population into multiple under-five and one five-and-over age group, and a simplified model of immune waning consisting of a one-step process characterized in terms of its rate and depth. Strain-specific parameters were used for the infection-to-paralysis ratio and the relative transmissibility of that strain's oral polio vaccine (OPV) relative to wild poliovirus (WPV), but strain interference was not directly included in the model, for the sake of simplicity. This analysis showed that all three of ongoing waning, a poorly transmissible vaccine, a high transmission setting, and an extended period of poor-but-existent vaccine coverage combined to create a potential for extended silent circulation which could continue beyond the three-year threshold that the Global Polio Eradication Initiative (GPEI) has set for declaring elimination of wild poliovirus from a country[87]. Among the implications of these results is that empirical experience from eradication campaigns in Latin America cannot be relied on to guarantee that current guidelines are adequate as indicators of when it is safe to stop OPV use. Because of this, more research is needed to determine how best to combine acute flacid paralysis surveillance (AFPS) and environmental surveillance (ES), both to better understand the underlying transmission dynamics, and to reliably detect silent circulation that threatens the success of global eradication efforts.

In Chapter V, originally the supplementary material to the paper whose contents are presented in Chapter IV, I analyzed several aspects of this model in greater depth. One aspect of these analyses was the formulation of an effective reproduction number (R_{eff}) analogue to the Next Generation Matrix basic reproduction number (R_0)[25]. Using this, I was able to show that the effective reproduction number for our

model could be decomposed into a simple sum of contributions from first infections of individuals under five years old, reinfections of individuals under five years old, first infections of individuals five and over, and reinfections of individuals five and over, thereby motivating and justifying a large portion of the analysis presented in Chapter IV. In addition, I analyzed in greater detail how the rate and depth of waning each contribute to the fraction of transmissions that come from reinfections vs. first infections at the endemic equilibrium. Finally, I compared the simplified model of waning that we used to the more detailed and realistic model presented by Wagner et al.[130], and showed that the latter generated at least as much ongoing waning as ours, and more in some respects. From this, I concluded that our results can reasonably be expected to be robust to the realistic relaxation of a number of our simplifying assumptions.

Silent circulation involves small numbers of infected individuals, and the events that mark its beginning and end are fundamentally stochastic in nature. For both of these reasons, a deterministic compartmental model (DCM) is not ideal for analyses in which characteristics of silent circulation are the primary outcomes. In Chapters IV and V, we dealt with this through some reasonable, but fundamentally *ad hoc* approximations. In Chapter VI, I improved on this approach by creating a hybrid model that used stochastic methods for the small numbers of individuals who are first-infected and reinfected with WPV during silent circulation, while retaining a deterministic approach to dealing with the much larger numbers of individuals in other states (i.e., fully susceptible, partially susceptible, fully immune, first-infected with OPV, and reinfected with OPV). Using this approach, I was able to show that the central conclusions of the previous chapters still held: Low vaccine transmissibility and ongoing waning both independently acted to increase the risk of sustained silent

circulation, and also to increase the risk of such circulation ending in case return. Moreover, there was a positive interaction between the two in this respect.

Additionally, because this hybrid approach was able to generate a range of possible outcomes for a given parameter set, instead of just a single deterministic outcome, I was able to show the existence of substantial regions of plausible parameter space where the median outcome was silent circulation lasting 3 years or less, but the probability of substantially longer silent circulation, often ending in case return, remained unacceptably high. The implication of this is that even our empirical experience from eradication efforts in very high-transmission settings, such as Northern India, is not sufficient to provide the necessary level of confidence in our predictions of outcomes in the remaining endemic countries upon which to base potentially high-risk policy decisions. This highlights the two distinct classes of sources of uncertainty regarding outcomes that can be relevant in situations like these: Uncertainty about parameters, and uncertainty about what outcome will occur even *conditional on* a given set parameters. Of these, only the former is readily examined using a purely deterministic model, but stochastic and hybrid models have the potential to explore both, and the interactions between the two.

In this dissertation, I have presented models and analyses that highlighted the consequences of realistically relaxing often-unexamined simplifying assumptions that are common in the transmission modeling of two very different pathogens, poliovirus and HIV. These analyses have shown that the inferences generated by commonly used models are not robust to such relaxation. They are not the last word on the particular issues they examine, much less the transmission dynamics of those pathogens as a whole, nor are they intended to be. They have numerous simplifying assumptions of their own, and more data is necessary to constrain their respective

plausible parameter spaces. But they represent informative and novel first steps toward building models of these transmission systems that can generate more robust inferences about key issues of practical importance for public health efforts aimed at control or eradicatation.

APPENDICES

APPENDIX A

Supplementary Material for Strong Influence of Behavioral Dynamics on the Ability of Testing and Treating HIV to Stop Transmission

A.1 Chapter preface

In Chapter III, I presented a simplified, deterministic model of episodic risk in men who have sex with men (MSM). That model included multiple risk groups, assortative mixing (conceptualized in terms of multiple mixing sites), and a simplified version of the natural history of HIV infection consisting of a short, highly-infectious acute phase and a longer, less-infectious chronic phase. In Chapter III (which originally constituted the main text of a paper to which this appendix constituted the supplementary material[58]), I extended that model to incorporate (a simplified version of) universal test and treat (UT&T), a strategy for control of the HIV epidemic that relies on encouraging the routine, voluntary HIV testing of at-risk individuals, with a goal of immediately starting antiretroviral therapy (ART) for those who test positive, in order to achieve viral suppression and thereby reduce or (effectively) eliminate their ability to generate secondary transmissions[36].

In this appendix, I present a deeper mathematical exploration of that model. I also present a more detailed description of and justification for some of the analyses performed in Chapter III. Finally, I examine the results of replacing the simplified

natural history of HIV infection used in Chapter III with a more detailed and realistic model, and show that the qualitative results obtained in Chapter III continue to hold.

A.2 Supplementary Figures

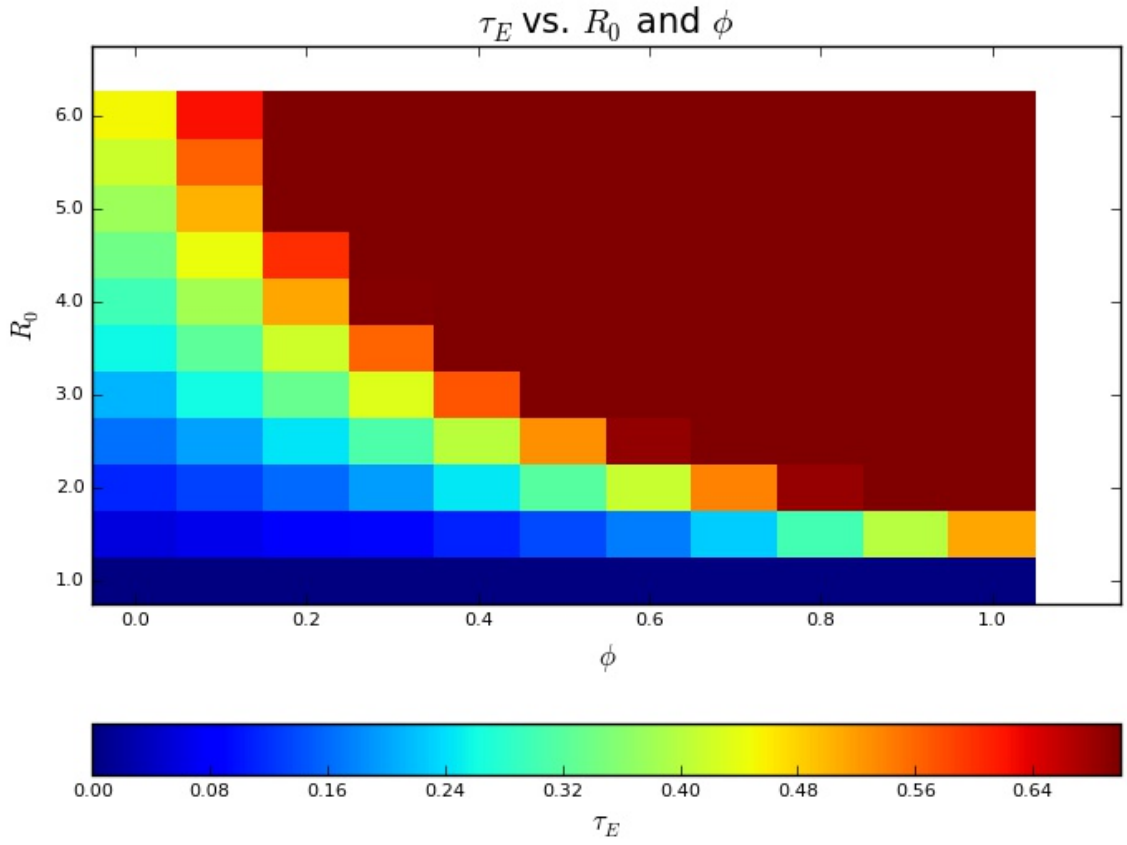


Figure A.1: Effective treatment rate required to achieve elimination of ongoing transmission (τ_E) vs. basic reproduction number (R_0) and fraction of transmissions from early infection (ϕ), under behavioral homogeneity. All values are 0 when $R_0 = 1$; and the maximum value obtained (when $R_0 = 6$ and $\phi = 1$) is 5.12.

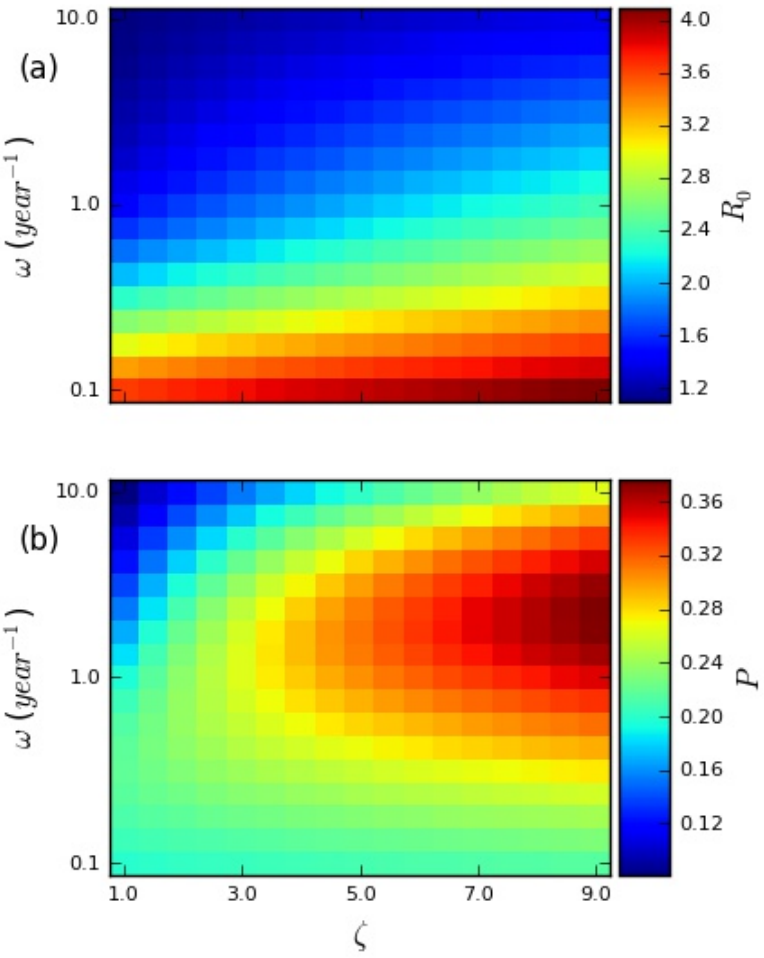


Figure A.2: Heatmaps plotting (a) the basic reproduction number (R_0), and (b) equilibrium prevalence (P) against the relative transmissibility during early HIV infection (ζ) and re-selection rate (ω), in the elaborated model. The average per-contact transmissibilities during EHI and chronic infection for each parameter set were chosen based on the constraints that (1) their ratio must be ζ and (2) the total transmission potential in a homogeneous system must be the same for all parameter sets. This transmission potential was chosen by setting the endemic prevalence for the lower-left parameter set to be 0.2. The other parameters used are summarized in Table A.1. Each panel has a scale that runs from the lowest to the highest value observed in that panel.

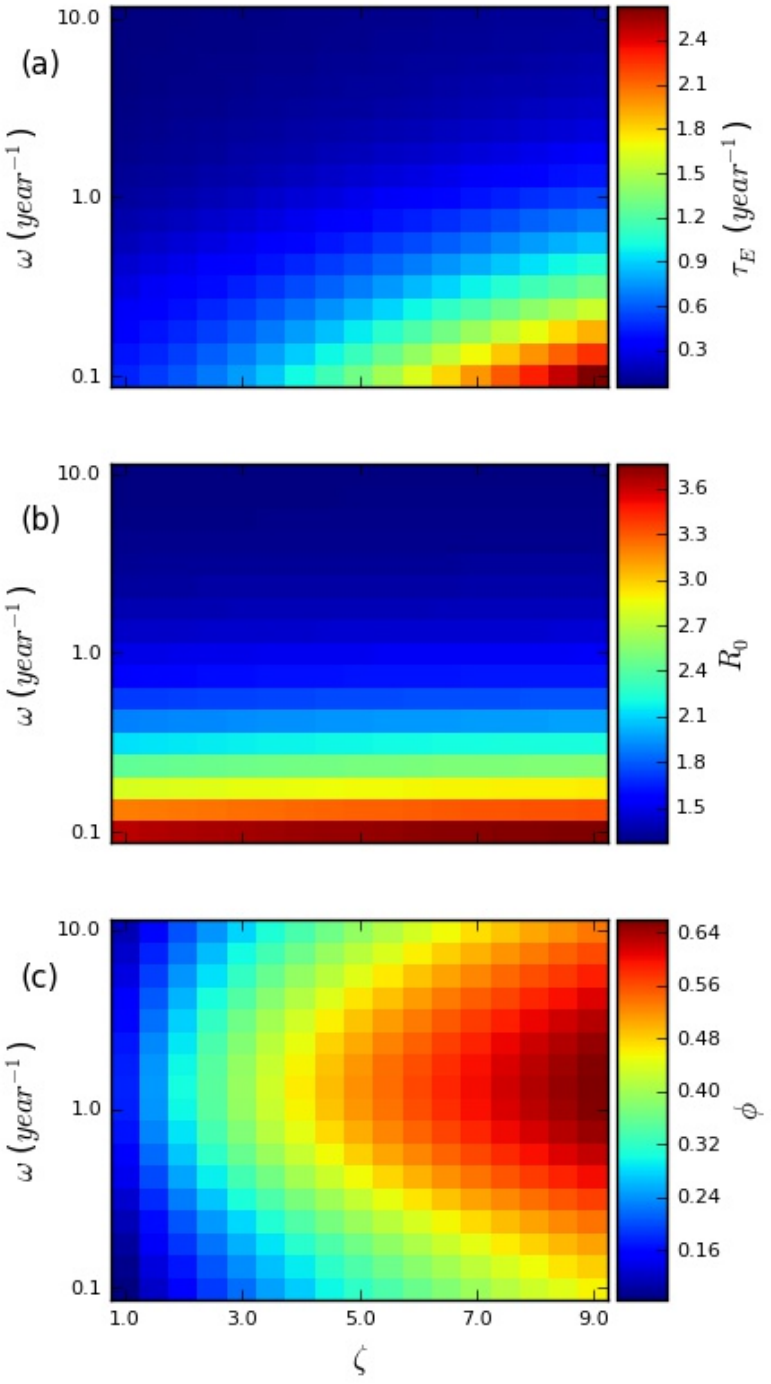


Figure A.3: Heatmaps plotting (a) the effective treatment rate required to achieve elimination (τ_E), (b) the basic reproduction number (R_0), and (c) the fraction of transmissions from early HIV infection (ϕ) against the relative transmissibility during early HIV infection (ζ) and re-selection rate (ω), in the elaborated system. The other parameters used are summarized in Table A.1. Each panel has a scale that runs from the lowest to the highest value observed in that panel.

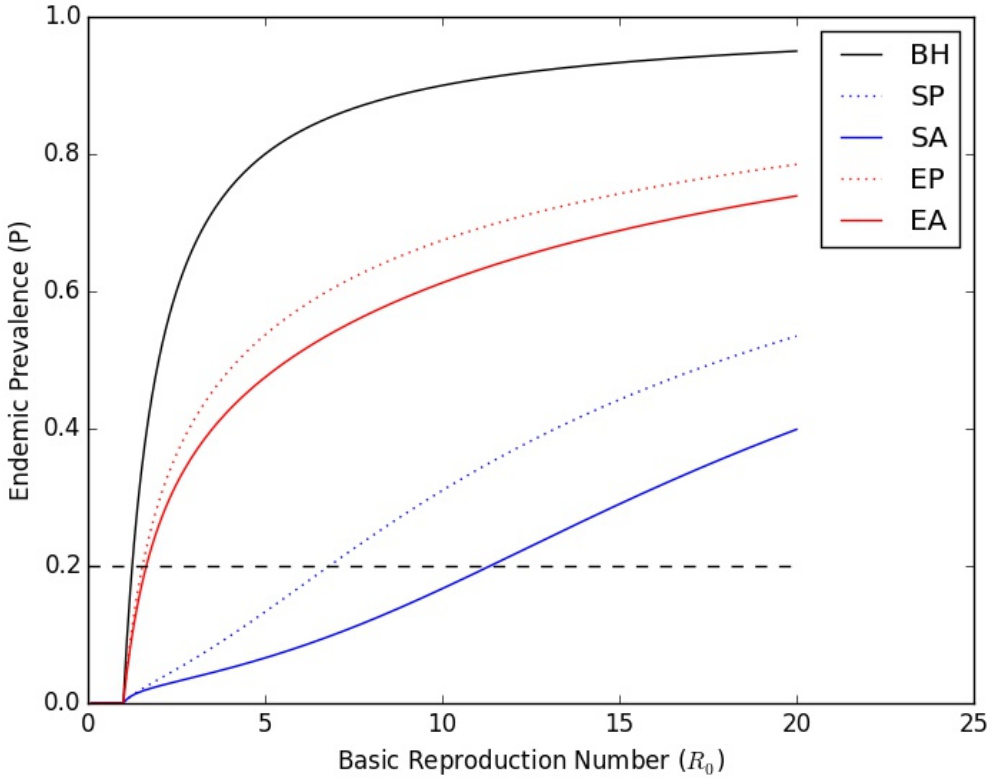


Figure A.4: Curves showing the prevalence as a function of R_0 when all parameters except the overall transmissibility are held constant, in the elaborated system. The curves shown are for the full (maximal) model (episodic risk, with assortative mixing (EA)), and several reduced models: Episodic risk, with proportional mixing (EP, the primary model in Chapter III and this appendix); Static risk heterogeneity, with assortative mixing (SA); Static risk heterogeneity, with proportional mixing (SP), and Behavioral Homogeneity (BH). The formulation of these reduced models is discussed in more detail in Section A.4.5. The dashed black line indicates a constant prevalence of 0.2, illustrating the drastically different values of R_0 that are possible when the prevalence is fixed.

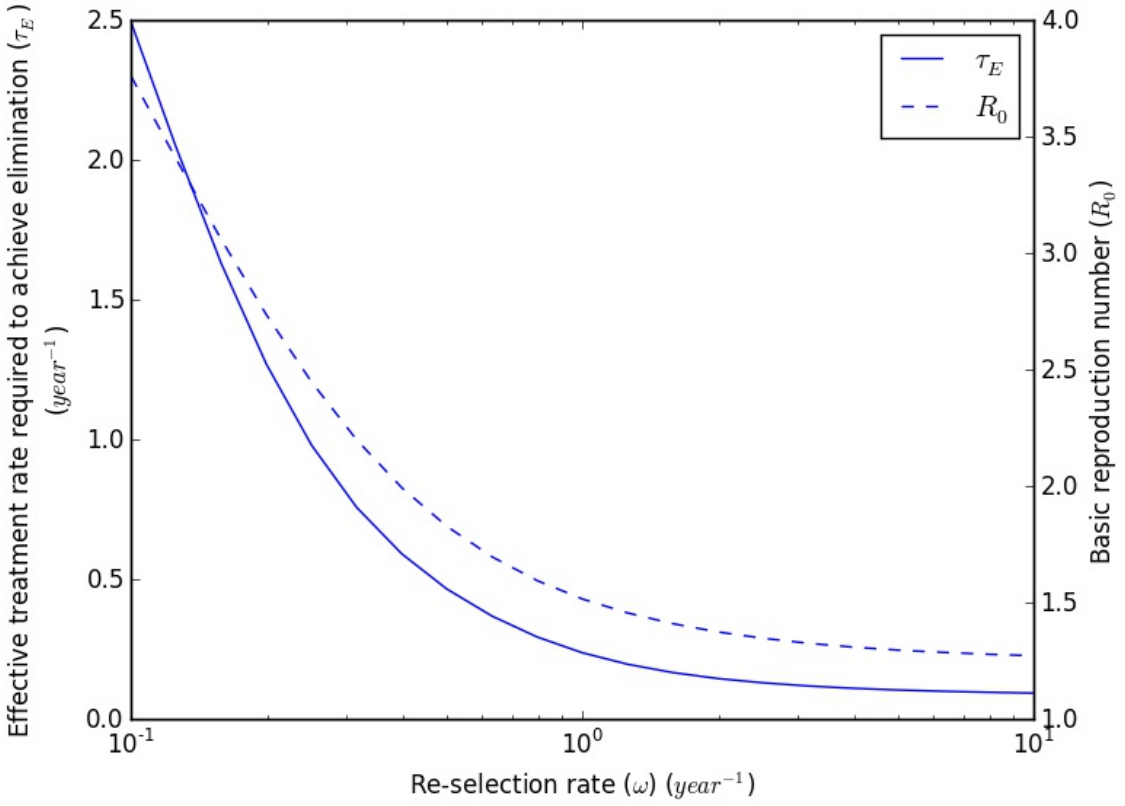


Figure A.5: The basic reproduction number (R_0) and effective treatment rate required to achieve elimination (τ_E) as a function of the re-selection rate (ω), in the elaborated system. The prevalence is fixed at 0.2, and the fraction of transmissions from EHI (ϕ) is fixed at 0.447[128]. To achieve this, the transmissibilities from each of acute and chronic infection are allowed to vary; all other parameters are fixed. Details are presented in Section A.4.7.

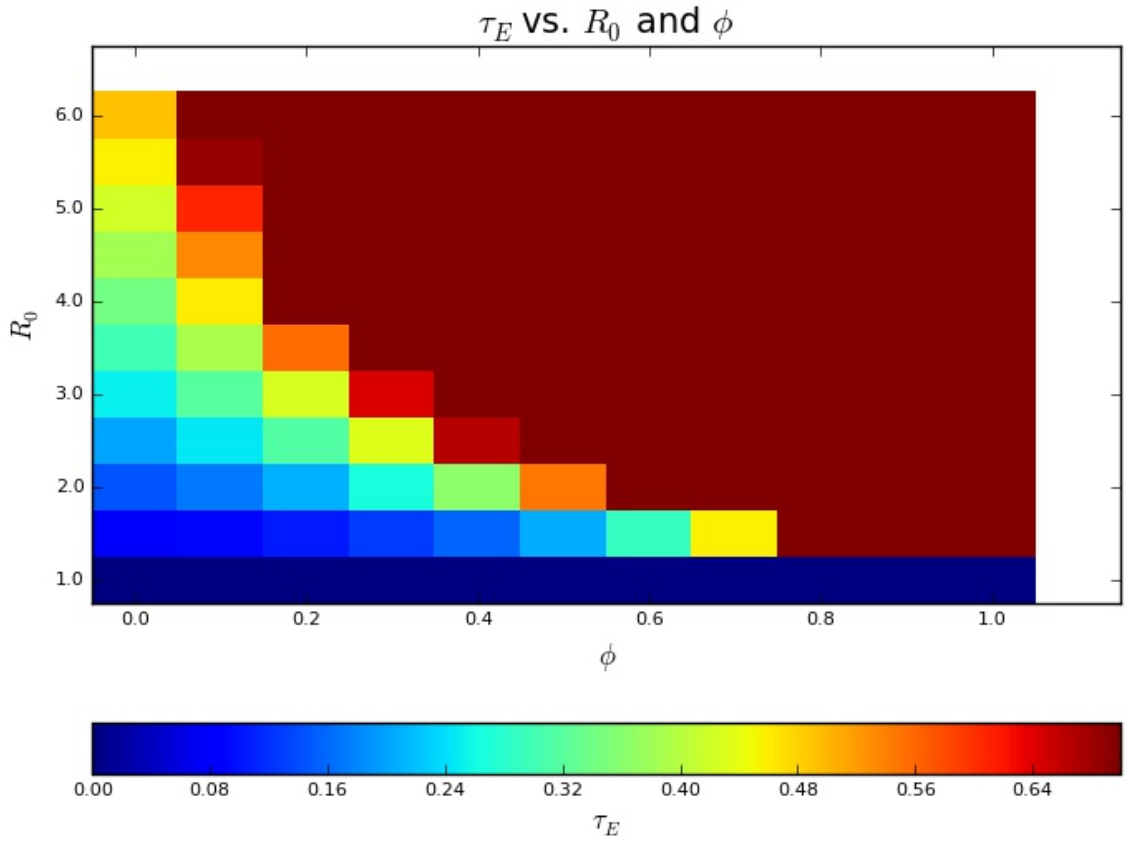


Figure A.6: Effective treatment rate required to achieve elimination of ongoing transmission (τ_E) vs. basic reproduction number (R_0) and fraction of transmissions from early infection (ϕ), under behavioral homogeneity, in the elaborated model. All values are 0 when $R_0 = 1$; and the maximum value obtained (when $R_0 = 6$ and $\phi = 1$) is 14.8.

A.3 Supplementary Tables

Symbol	Unit	Value(s)	Meaning
r_{HL}	-	21	Contact rate ratio between higher risk and lower risk individuals
f_{H}	-	0.1	Fraction of the population that is at higher risk at the disease-free equilibrium.
m	-	0	Fraction of an individual's contacts that are reserved for members of the same risk group. Throughout Chapter III, it is 0 unless stated otherwise.
τ	1/year	0	Effective treatment rate
ω	1/year	1	Rate at which we randomly re-select which risk group an individual is in (re-selection rate)
ζ	-	10	Relative transmissibility during EHI

Table A.1: Default parameters. All figures and results in Chapter III and this appendix use this parameter set, except as explicitly noted. Only parameters with "variable" values in Table 3.1 are shown here.

Symbol	Value(s)	Meaning
ν_1	$\mu + \gamma_1 + \tau$	Total rate of leaving the early-infected and untreated state.
ν_2	$\mu + \gamma_2 + \tau$	Total rate of leaving the chronically infected and untreated state.
g_{H}	$\frac{f_{\mathrm{H}} r_{\mathrm{HL}}}{1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)}$	Fraction of all contacts that are made by higher risk individuals at the disease-free equilibrium.
g_{L}	$\frac{f_{\mathrm{L}}}{1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)}$	Fraction of all contacts that are made by lower risk individuals at the disease-free equilibrium.
$c_V(\chi)^2$	$\frac{f_{\mathrm{H}} f_{\mathrm{L}} (r_{\mathrm{HL}} - 1)^2}{(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1))^2}$	The variance of the contact rate divided by the square of the mean contact rate (i.e. the coefficient of variation squared).
r	$\frac{m r_{\mathrm{HL}}}{(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1))^2}$	Scaled assortativity, used in calculating R_0.
H_1	$\frac{\chi \beta_1}{\nu_1}$	The contribution of transmissions during EHI to R_0 in the corresponding behaviorally homogeneous system.
H_2	$\frac{\chi \gamma_1 \beta_2}{\nu_1 \nu_2}$	The contribution of transmissions during chronic infection to R_0 in the corresponding behaviorally homogeneous system.
ψ_1	$\frac{\nu_1}{\nu_1 + \omega}$	The fraction of the average time spent in untreated EHI that occurs prior to the first re-selection of that individual's contact rate after his infection.
ψ_2	$\frac{\nu_1 \nu_2}{(\nu_1 + \omega)(\nu_2 + \omega)}$	The fraction of the average time spent in untreated chronic infection that occurs prior to the first re-selection of that individual's contact rate after his infection.
H	$H_1 + H_2$	The contribution of all transmissions to R_0 in the corresponding behaviorally homogeneous system (i.e. R_0 for the corresponding behaviorally homogeneous system).
ψ	$\frac{\psi_1 H_1 + \psi_2 H_2}{H_1 + H_2}$	The fraction of the average number of secondary transmissions per index case, during exponential growth and under behavioral homogeneity, that occur prior to the first re-selection of that individual's contact rate after his infection.

Table A.2: Derived parameters used in the calculation and presentation of R_0.

Stage Number	Grouped Under	Mean Duration ($1/\gamma_i$)	Per-contact transmission rate (β_i)
1	Acute	1 week	0.003
2	Acute	1 week	0.03
3	Acute	1 week	0.04
4	Acute	1 week	0.03
5	Acute	16.6 weeks	0.02
6	Chronic	1.9 years	0.0007
7	Chronic	1.9 years	0.0007
8	Chronic	1.9 years	0.0007
9	Chronic ("Early AIDS")	0.9 years	0.006

Table A.3: Stage of infection parameters for the elaborated model

A.4 Supplementary Methods

A.4.1 Detailed description of the underlying transmission model

Our model of transmission dynamics is a deterministic compartmental model (DCM), but, as discussed in Chapter III, we will frequently refer to "individuals" to describe the meaning of various flows, using the fact that a DCM is the limiting case of an individual-based model in an infinitely large, thoroughly-mixing population. For convenience, the total population size is set equal to 1 at the disease free equilibrium. Consequently, the "birth" (entry into the population of sexually active MSM) rate is equal to the natural "mortality" (departure from the sexually active population) rate μ. For simplicity, we assume that this birth rate does not depend on the size of the current population of MSM. Throughout both Chapter III and this appendix, μ is fixed at 0.0248/year. There is no age structure in our model. We also exclude importation of HIV from outside the population from our model.

Individuals are divided into two risk groups, differentiated by their (sexual) contact rates, which we denote χ_{H} for the higher risk group and χ_{L} for the lower risk group. We parameterize this division by the fraction of individuals who enter the population in the higher risk group (f_{H}), and the ratio of contact rates between the two groups (r_{HL}) The mean contact rate χ is fixed at 20/year. All contacts are assumed to be instantaneous and symmetric. Individuals in each risk group reserve a fraction m of their contacts for members of their own risk group (0 by default), with all other contacts being made proportionately.

Individuals do not necessarily remain in the same risk group for the full duration of their stay in the sexually active population; we re-select which risk group they are in at a rate ω, which we refer to as the "re-selection rate." When we do so, we assign them to each risk group with the same probability as when they first entered the

sexually active population; consequently, f_{H} is also the fraction of individuals who are in the higher risk group at the disease-free equilibrium.

We simplify the natural history of infection to a relatively brief early phase (early HIV infection, EHI) followed by a longer chronic phase. During early infection, the per-act transmission probability is β_1, and during chronic infection, it falls to β_2. β_2 is in general allowed to vary in order to attain a desired endemic prevalence or basic reproduction number; β_1 is the product of β_2 and the relative transmissibility during EHI (relative to chronic infection), which we denote ζ. Progression from early to chronic infection occurs at a rate of γ_1 (1/year throughout Chapter III and this appendix), regardless of treatment status, and "progression" from chronic infection to death or departure from the sexually active population due to HIV infection occurs at a rate of γ_2 (1/8.45 years) if untreated, and does not occur if treated ($\gamma_{2\mathrm{T}} = 0$) – note, however, that none of the calculations or results in Chapter III or in this appendix depend on the value of $\gamma_{2\mathrm{T}}$.

The treatment cascade is simplified to a binary distinction between individuals who are "effectively treated" (those who have attained sustained viral suppression) and individuals who are not (everyone else). Transition from the latter group to the former occurs at an "effective treatment rate" of τ. To simplify the model, we do not include virologic failure after viral suppression, and we assume that virally suppressed individuals are completely non-transmitting.

The above-described dynamics are summarized in the following set of linked dif-

ferential equations (symbols denoting subpopulations are defined in Table 3.1):

$$\begin{aligned}
\frac{d}{dt}(S_{\mathrm{L}}) &= f_{\mathrm{L}}\mu - (\mu + \chi_{\mathrm{L}}\lambda_{\mathrm{L}} + f_{\mathrm{H}}\omega)S_{\mathrm{L}} + f_{\mathrm{L}}\omega S_{\mathrm{H}} \\
\frac{d}{dt}(S_{\mathrm{H}}) &= f_{\mathrm{H}}\mu - (\mu + \chi_{\mathrm{H}}\lambda_{\mathrm{H}} + f_{\mathrm{L}}\omega)S_{\mathrm{H}} + f_{\mathrm{H}}\omega S_{\mathrm{L}} \\
\frac{d}{dt}(A_{\mathrm{L,U}}) &= \chi_{\mathrm{L}}\lambda_{\mathrm{L}}S_{\mathrm{L}} - (\mu + \gamma_1 + f_{\mathrm{H}}\omega + \tau)A_{\mathrm{L,U}} + f_{\mathrm{L}}\omega A_{\mathrm{H,U}} \\
\frac{d}{dt}(A_{\mathrm{H,U}}) &= \chi_{\mathrm{H}}\lambda_{\mathrm{H}}S_{\mathrm{H}} - (\mu + \gamma_1 + f_{\mathrm{L}}\omega + \tau)A_{\mathrm{H,U}} + f_{\mathrm{H}}\omega A_{\mathrm{L,U}} \\
\frac{d}{dt}(C_{\mathrm{L,U}}) &= \gamma_1 A_{\mathrm{L,U}} - (\mu + \gamma_2 + f_{\mathrm{H}}\omega + \tau)C_{\mathrm{L,U}} + f_{\mathrm{L}}\omega C_{\mathrm{H,U}} \\
\frac{d}{dt}(C_{\mathrm{H,U}}) &= \gamma_1 A_{\mathrm{H,U}} - (\mu + \gamma_2 + f_{\mathrm{L}}\omega + \tau)C_{\mathrm{H,U}} + f_{\mathrm{H}}\omega C_{\mathrm{L,U}} \\
\frac{d}{dt}(A_{\mathrm{L,T}}) &= \tau A_{\mathrm{L,U}} - (\mu + \gamma_1 + f_{\mathrm{H}}\omega)A_{\mathrm{L,T}} + f_{\mathrm{L}}\omega A_{\mathrm{H,T}} \\
\frac{d}{dt}(A_{\mathrm{H,T}}) &= \tau A_{\mathrm{H,U}} - (\mu + \gamma_1 + f_{\mathrm{L}}\omega)A_{\mathrm{H,T}} + f_{\mathrm{H}}\omega A_{\mathrm{L,T}} \\
\frac{d}{dt}(C_{\mathrm{L,T}}) &= \gamma_1 A_{\mathrm{L,T}} + \tau C_{\mathrm{L,U}} - (\mu + \gamma_{2\mathrm{T}} + f_{\mathrm{H}}\omega)C_{\mathrm{L,T}} + f_{\mathrm{L}}\omega C_{\mathrm{H,T}} \\
\frac{d}{dt}(C_{\mathrm{H,T}}) &= \gamma_1 A_{\mathrm{H,T}} + \tau C_{\mathrm{H,U}} - (\mu + \gamma_{2\mathrm{T}} + f_{\mathrm{L}}\omega)C_{\mathrm{H,T}} + f_{\mathrm{H}}\omega C_{\mathrm{L,T}}
\end{aligned}$$

where

$$\begin{aligned}
\lambda_{\mathrm{L}} &= m\frac{\beta_1 A_{\mathrm{L,U}} + \beta_2 C_{\mathrm{L,U}}}{L} + (1-m)\frac{\beta_1(\chi_{\mathrm{L}}A_{\mathrm{L,U}} + \chi_{\mathrm{H}}A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}}C_{\mathrm{L,U}} + \chi_{\mathrm{H}}C_{\mathrm{H,U}})}{\chi_{\mathrm{L}}L + \chi_{\mathrm{H}}H} \\
\lambda_{\mathrm{H}} &= m\frac{\beta_1 A_{\mathrm{H,U}} + \beta_2 C_{\mathrm{H,U}}}{H} + (1-m)\frac{\beta_1(\chi_{\mathrm{L}}A_{\mathrm{L,U}} + \chi_{\mathrm{H}}A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}}C_{\mathrm{L,U}} + \chi_{\mathrm{H}}C_{\mathrm{H,U}})}{\chi_{\mathrm{L}}L + \chi_{\mathrm{H}}H}
\end{aligned}$$

A.4.2 Calculation of transmission potentials

The rate of leaving untreated EHI is simply the sum of the rates of dying, progressing to chronic infection, and becoming treated: $\nu_1 = \mu + \gamma_1 + \tau$. Therefore, the average time spent in untreated EHI is $1/\nu_1$. During exponential growth, and under behavioral homogeneity, the transmission rate during this time is $\chi\beta_1$. Consequently, the expected number of transmissions during EHI per index case (the transmission

potential from EHI) is

$$H_1 = \frac{\chi\beta_1}{\nu_1}. \tag{A.1}$$

Similarly, the expected number of transmissions during chronic infection per index case that reaches chronic infection untreated is $\chi\beta/\nu_2$, where $\nu_2 = \mu + \gamma_2 + \tau$. However, only a fraction γ_1/ν_1 of index cases reach chronic infection before dying or becoming treated. Consequently, the expected number of transmissions during chronic infection per index case (the transmission potential from chronic infection) is

$$H_2 = \frac{\chi\gamma_1\beta_2}{\nu_1\nu_2}. \tag{A.2}$$

The total transmission potential is simply the sum of the transmission potentials from EHI and from chronic infection:

$$H = H_1 + H_2. \tag{A.3}$$

A.4.3 Calculation of remaining heterogeneity effects

Consider the subpopulation of individuals who are acutely infected and untreated, and whose contact rate has not been re-selected since they became infected. Individuals leave this subpopulation at a rate of ν_1, and we re-select their contact rates at a rate of ω. Consequently, the fraction of acutely infected, untreated individuals whose contact rates we have not re-selected since their infection when they cease to be both acutely infected and untreated is

$$\psi_1 = \frac{\nu_1}{\nu_1 + \omega}. \tag{A.4}$$

Because the dynamics of this system are all first order, this is also the (average) fraction of time spent acutely infected and untreated that occurs prior to the first

post-infection re-selection of an individual's contact rate. Similarly, the fraction of time spent chronically infected and untreated *by individuals who enter chronic infection prior to the first post-infection re-selection of their contact rates* that occurs prior to the first post-infection re-selection of their contact rates is $\nu_2/(\nu_2 + \omega)$. However, only ψ_1 of individuals entering chronic infection untreated do so without having already had their contact rates re-selected. Consequently, the fraction of time spent chronically infected and untreated by all individuals that occurs prior to the first post-infection re-selection of their contact rates is

$$\psi_2 = \psi_1\left(\frac{\nu_2}{(\nu_2+\omega)}\right) = \frac{\nu_1\nu_2}{(\nu_1+\omega)(\nu_2+\omega)}. \tag{A.5}$$

Because ψ represents the fraction of secondary transmissions (during exponential growth and under behavioral homogeneity) that occur prior to the first post-infection re-selection of the transmitter's contact rate, its numerator is the sum of the number of such transmissions during EHI and the number of such transmission during chronic infection, while its denominator is the sum of all transmissions during EHI and all transmissions during chronic infection:

$$\psi = \frac{\psi_1 H_1 + \psi_2 H_2}{H_1 + H_2}. \tag{A.6}$$

A.4.4 Construction of Figure 3.1

We used the base parameters detailed in Table A.1, except for ζ and ω, which we varied, with ranges of 1–9 and 0.1–10 (the latter on a logarithmic scale), respectively. To begin, we found by numerical simulation values for β_1 and β_2 (related by the equation $\beta_1 = \zeta\beta_2$) that would result in an endemic prevalence of 0.2 when $\zeta = 1$ and $\omega = 0.1$, corresponding to the lower left corner of each panel of Figure 3.1. We then calculated the total (homogeneous) transmission potential H, using the formulas

in Table A.2. For each of the remaining parameter sets, we began by setting ζ and ω for that parameter set, leaving the other parameters unchanged, and calculating the new total transmission potential H'. Because the total transmission potential is proportional to β_2 when all other parameters are fixed (including ζ, but not including the derived parameter β_1), and we wanted all parameter sets to have the same total transmission potential, we then multiplied β_2 by the ratio H/H'.

For each of the parameter sets we found in this fashion, we found the endemic prevalence by numerical simulation and calculated R_0 using the methods described in Section A.4.6. These results were then plotted as heatmaps using the python function matplotlib.pyplot.pcolor. All code used can be obtained by the means discussed in Appendix B.

A.4.5 Construction of submodels for Figure 3.3

The standard model (episodic risk, proportional mixing (EP)) uses the full parameter set given in Table A.1. For all of the other models, one or more parameters are changed: ω, indicating the degree to which risk heterogeneity is episodic, is set to 0 for the models SA (static risk heterogeneity, assortative mixing), SP (static risk heterogeneity, proportional mixing), and BH (behavioral homogeneity). m, indicating the degree to which mixing is assortative, is set to 0.5 in the models EA (episodic risk, assortative mixing) and SA. r_{HL}, indicating the degree of risk heterogeneity, is set to 0 in model BH only. Although model BH still technically contains two behavioral groups, they are now indistinguishable from each other in all respects, and are therefore effectively the same group.

A.4.6 Calculation of outcomes given a parameter set

Our method for finding the basic reproduction number R_0 is that developed by Diekmann et al.[25] We will briefly review that method in the course of describing our use of it.

Given a set of differential equations for a transmission system, we first construct a linearization of the equations for infectious compartments only, at the disease-free equilibrium (DFE). For the differential equations in Section A.4.1, at the disease-free equilibrium, $S_{\mathrm{H}} = f_{\mathrm{H}}$ and $S_{\mathrm{L}} = f_{\mathrm{L}}$. By substituting these values into the differential equations in Section A.4.1, and noting that $\chi = f_{\mathrm{H}}\chi_{\mathrm{H}} + f_{\mathrm{L}}\chi_{\mathrm{L}}$ (Table 3.2), we obtain the following linearization:

$$\begin{aligned}
\frac{d}{dt}(A_{\mathrm{H,U}}) &= \chi_{\mathrm{H}}\lambda_{\mathrm{H}}f_{\mathrm{H}} - (\nu_1 + f_{\mathrm{L}}\omega)A_{\mathrm{H,U}} + f_{\mathrm{H}}\omega A_{\mathrm{L,U}} \\
\frac{d}{dt}(A_{\mathrm{L,U}}) &= \chi_{\mathrm{L}}\lambda_{\mathrm{L}}f_{\mathrm{L}} - (\nu_1 + f_{\mathrm{H}}\omega)A_{\mathrm{L,U}} + f_{\mathrm{L}}\omega A_{\mathrm{H,U}} \\
\frac{d}{dt}(C_{\mathrm{H,U}}) &= \gamma_1 A_{\mathrm{H,U}} - (\nu_2 + f_{\mathrm{L}}\omega)C_{\mathrm{H,U}} + f_{\mathrm{H}}\omega C_{\mathrm{L,U}} \\
\frac{d}{dt}(C_{\mathrm{L,U}}) &= \gamma_1 A_{\mathrm{L,U}} - (\nu_2 + f_{\mathrm{H}}\omega)C_{\mathrm{L,U}} + f_{\mathrm{L}}\omega C_{\mathrm{H,U}}
\end{aligned}$$

where

$$\begin{aligned}
\lambda_{\mathrm{H}} &= m\frac{\beta_1 A_{\mathrm{H,U}} + \beta_2 C_{\mathrm{H,U}}}{f_{\mathrm{H}}} + (1-m)\frac{\beta_1(\chi_{\mathrm{L}}A_{\mathrm{L,U}} + \chi_{\mathrm{H}}A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}}C_{\mathrm{L,U}} + \chi_{\mathrm{H}}C_{\mathrm{H,U}})}{\chi} \\
\lambda_{\mathrm{L}} &= m\frac{\beta_1 A_{\mathrm{L,U}} + \beta_2 C_{\mathrm{L,U}}}{f_{\mathrm{L}}} + (1-m)\frac{\beta_1(\chi_{\mathrm{L}}A_{\mathrm{L,U}} + \chi_{\mathrm{H}}A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}}C_{\mathrm{L,U}} + \chi_{\mathrm{H}}C_{\mathrm{H,U}})}{\chi}
\end{aligned}$$

Because we will need to distinguish between terms representing transitions of infected individuals between compartments or out of the system and terms representing new

transmissions, let us denote these as $\{x_i\}$ and $\{y_i\}$, respectively:

$$\begin{aligned}
\frac{d}{dt}(A_{\mathrm{H,U}}) &= x_{\mathrm{A,H}} + y_{\mathrm{A,H}} \\
\frac{d}{dt}(A_{\mathrm{L,U}}) &= x_{\mathrm{A,L}} + y_{\mathrm{A,L}} \\
\frac{d}{dt}(C_{\mathrm{H,U}}) &= x_{\mathrm{C,H}} + y_{\mathrm{C,H}} \\
\frac{d}{dt}(C_{\mathrm{L,U}}) &= x_{\mathrm{C,L}} + y_{\mathrm{C,L}}
\end{aligned}$$

where

$$\begin{aligned}
x_{\mathrm{A,H}} =& -(\nu_1 + f_{\mathrm{L}}\omega)A_{\mathrm{H,U}} + f_{\mathrm{H}}\omega A_{\mathrm{L,U}} \\
x_{\mathrm{A,L}} =& -(\nu_1 + f_{\mathrm{H}}\omega)A_{\mathrm{L,U}} + f_{\mathrm{L}}\omega A_{\mathrm{H,U}} \\
x_{\mathrm{C,H}} =& \gamma_1 A_{\mathrm{H,U}} - (\nu_2 + f_{\mathrm{L}}\omega)C_{\mathrm{H,U}} + f_{\mathrm{H}}\omega C_{\mathrm{L,U}} \\
x_{\mathrm{C,L}} =& \gamma_1 A_{\mathrm{L,U}} - (\nu_2 + f_{\mathrm{H}}\omega)C_{\mathrm{L,U}} + f_{\mathrm{L}}\omega C_{\mathrm{H,U}} \\
y_{\mathrm{A,H}} =& \chi_{\mathrm{H}}\lambda_{\mathrm{H}} f_{\mathrm{H}} \\
=& m\chi_{\mathrm{H}}(\beta_1 A_{\mathrm{H,U}} + \beta_2 C_{\mathrm{H,U}}) \\
&+ (1-m)\chi_{\mathrm{H}} f_{\mathrm{H}} \frac{\beta_1(\chi_{\mathrm{L}} A_{\mathrm{L,U}} + \chi_{\mathrm{H}} A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}} C_{\mathrm{L,U}} + \chi_{\mathrm{H}} C_{\mathrm{H,U}})}{\chi} \\
y_{\mathrm{A,L}} =& \chi_{\mathrm{L}}\lambda_{\mathrm{L}} f_{\mathrm{L}} \\
=& m\chi_{\mathrm{L}}(\beta_1 A_{\mathrm{L,U}} + \beta_2 C_{\mathrm{L,U}}) \\
&+ (1-m)\chi_{\mathrm{L}} f_{\mathrm{L}} \frac{\beta_1(\chi_{\mathrm{L}} A_{\mathrm{L,U}} + \chi_{\mathrm{H}} A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}} C_{\mathrm{L,U}} + \chi_{\mathrm{H}} C_{\mathrm{H,U}})}{\chi} \\
y_{\mathrm{C,H}} =& y_{\mathrm{C,L}} = 0
\end{aligned}$$

We can simplify the transmission terms a little more by introducing the symbols g_{L} and g_{H} to represent the fraction of all contacts at the disease-free equilibrium that

are made by low-risk and high-risk individuals, respectively:

$$\begin{aligned} g_{\mathrm{L}} =& \frac{f_{\mathrm{L}}\chi_{\mathrm{L}}}{f_{\mathrm{L}}\chi_{\mathrm{L}} + f_{\mathrm{H}}\chi_{\mathrm{H}}} = \frac{f_{\mathrm{L}}\chi_{\mathrm{L}}}{\chi} = \frac{f_{\mathrm{L}}}{1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)} \\ g_{\mathrm{H}} =& \frac{f_{\mathrm{H}}\chi_{\mathrm{H}}}{f_{\mathrm{L}}\chi_{\mathrm{L}} + f_{\mathrm{H}}\chi_{\mathrm{H}}} = \frac{f_{\mathrm{H}}\chi_{\mathrm{H}}}{\chi} = \frac{f_{\mathrm{H}}r_{\mathrm{HL}}}{1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)} \end{aligned}$$

This produces the rather more elegant equations:

$$\begin{aligned} y_{\mathrm{A,H}} =& m\chi_{\mathrm{H}}(\beta_1 A_{\mathrm{H,U}} + \beta_2 C_{\mathrm{H,U}}) \\ &+ (1-m)g_{\mathrm{H}}(\beta_1(\chi_{\mathrm{L}}A_{\mathrm{L,U}} + \chi_{\mathrm{H}}A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}}C_{\mathrm{L,U}} + \chi_{\mathrm{H}}C_{\mathrm{H,U}})) \\ =& \chi_{\mathrm{H}}(g_{\mathrm{H}} + mg_{\mathrm{L}})(\beta_1 A_{\mathrm{H,U}} + \beta_2 C_{\mathrm{H,U}}) + \chi_{\mathrm{L}}(1-m)g_{\mathrm{L}}(\beta_1 A_{\mathrm{L,U}} + \beta_2 C_{\mathrm{L,U}}) \\ y_{\mathrm{A,L}} =& m\chi_{\mathrm{L}}(\beta_1 A_{\mathrm{L,U}} + \beta_2 C_{\mathrm{L,U}}) \\ &+ (1-m)g_{\mathrm{L}}(\beta_1(\chi_{\mathrm{L}}A_{\mathrm{L,U}} + \chi_{\mathrm{H}}A_{\mathrm{H,U}}) + \beta_2(\chi_{\mathrm{L}}C_{\mathrm{L,U}} + \chi_{\mathrm{H}}C_{\mathrm{H,U}})) \\ =& \chi_{\mathrm{L}}(g_{\mathrm{L}} + mg_{\mathrm{H}})(\beta_1 A_{\mathrm{L,U}} + \beta_2 C_{\mathrm{L,U}}) + \chi_{\mathrm{H}}(1-m)g_{\mathrm{H}}(\beta_1 A_{\mathrm{H,U}} + \beta_2 C_{\mathrm{H,U}}) \end{aligned}$$

From these, we construct the Jacobian matrix J:

$$J = \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) & \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) \\ \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) & \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) \\ \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) & \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) \\ \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) & \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) \end{pmatrix}$$

This and subsequent matrices can made both less cluttered and easier to understand by presenting them as *block matrices*, i.e., as matrices composed of submatrices. For this particular model, both the most intuitive and the most mathematically convenient way of doing this is by grouping together the two acutely-infected subpopulations on the one hand, and the two chronically infected subpopulations on the

other:

$$J = \begin{pmatrix} J_{\mathrm{A,A}} & J_{\mathrm{A,C}} \\ J_{\mathrm{C,A}} & J_{\mathrm{C,C}} \end{pmatrix}$$

where

$$J_{\mathrm{A,A}} = \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) \\ \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{A,H}} + y_{\mathrm{A,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{A,H}} + y_{\mathrm{A,H}}) \\ \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{L,H}} + y_{\mathrm{L,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{L,H}} + y_{\mathrm{L,H}}) \end{pmatrix}$$

$$J_{\mathrm{A,C}} = \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{H,U}})\right) \\ \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(A_{\mathrm{L,U}})\right) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{A,H}} + y_{\mathrm{A,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{A,H}} + y_{\mathrm{A,H}}) \\ \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{L,H}} + y_{\mathrm{L,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{L,H}} + y_{\mathrm{L,H}}) \end{pmatrix}$$

$$J_{\mathrm{C,A}} = \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) \\ \frac{d}{dA_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) & \frac{d}{dA_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{C,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{C,H}}) \\ \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{C,L}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{C,L}}) \end{pmatrix}$$

$$J_{\mathrm{C,C}} = \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{H,U}})\right) \\ \frac{d}{dC_{\mathrm{H,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) & \frac{d}{dC_{\mathrm{L,U}}}\left(\frac{d}{dt}(C_{\mathrm{L,U}})\right) \end{pmatrix}$$

$$= \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{C,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{C,H}}) \\ \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{C,L}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{C,L}}) \end{pmatrix}$$

We then decompose J into the sum of a transition matrix Σ and a transmission matrix T, using the distinction that we have already made between terms represent-

ing transmitions and terms representing transmissions:

$$J = T + \Sigma = \begin{pmatrix} T_{\mathrm{A,A}} & T_{\mathrm{A,C}} \\ T_{\mathrm{C,A}} & T_{\mathrm{C,C}} \end{pmatrix} + \begin{pmatrix} \Sigma_{\mathrm{A,A}} & \Sigma_{\mathrm{A,C}} \\ \Sigma_{\mathrm{C,A}} & \Sigma_{\mathrm{C,C}} \end{pmatrix}$$

where

$$\Sigma_{\mathrm{A,A}} = \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{A,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{A,H}}) \\ \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{L,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{L,H}}) \end{pmatrix} = \begin{pmatrix} -(\nu_1 + f_{\mathrm{L}}\omega) & f_{\mathrm{H}}\omega \\ f_{\mathrm{L}}\omega & -(\nu_1 + f_{\mathrm{H}}\omega) \end{pmatrix}$$

$$\Sigma_{\mathrm{A,C}} = \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{A,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{A,H}}) \\ \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{L,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{L,H}}) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \mathbf{0}_{2,2}$$

$$\Sigma_{\mathrm{C,A}} = \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{C,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{C,H}}) \\ \frac{d}{dA_{\mathrm{H,U}}}(x_{\mathrm{C,L}}) & \frac{d}{dA_{\mathrm{L,U}}}(x_{\mathrm{C,L}}) \end{pmatrix} = \begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_1 \end{pmatrix} = \gamma_1 \mathbf{I}_2$$

$$\Sigma_{\mathrm{C,C}} = \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{C,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{C,H}}) \\ \frac{d}{dC_{\mathrm{H,U}}}(x_{\mathrm{C,L}}) & \frac{d}{dC_{\mathrm{L,U}}}(x_{\mathrm{C,L}}) \end{pmatrix} = \begin{pmatrix} -(\nu_2 + f_{\mathrm{L}}\omega) & f_{\mathrm{H}}\omega \\ f_{\mathrm{L}}\omega & -(\nu_2 + f_{\mathrm{H}}\omega) \end{pmatrix}$$

$$T_{\mathrm{A,A}} = \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}(y_{\mathrm{A,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(y_{\mathrm{A,H}}) \\ \frac{d}{dA_{\mathrm{H,U}}}(y_{\mathrm{L,H}}) & \frac{d}{dA_{\mathrm{L,U}}}(y_{\mathrm{L,H}}) \end{pmatrix} = \begin{pmatrix} \chi_{\mathrm{H}}\beta_1(g_{\mathrm{H}} + mg_{\mathrm{L}}) & \chi_{\mathrm{L}}\beta_1(1-m)g_{\mathrm{H}} \\ \chi_{\mathrm{H}}\beta_1(1-m)g_{\mathrm{L}} & \chi_{\mathrm{L}}\beta_1(g_{\mathrm{L}} + mg_{\mathrm{H}}) \end{pmatrix}$$

$$= \frac{\chi\beta_1}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \left(\begin{pmatrix} f_{\mathrm{H}} r_{\mathrm{HL}}^2 & f_{\mathrm{H}} r_{\mathrm{HL}} \\ f_{\mathrm{L}} r_{\mathrm{HL}} & f_{\mathrm{L}} \end{pmatrix} + m r_{\mathrm{HL}} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \right)$$

$$T_{\mathrm{A,C}} = \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}(y_{\mathrm{A,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(y_{\mathrm{A,H}}) \\ \frac{d}{dC_{\mathrm{H,U}}}(y_{\mathrm{L,H}}) & \frac{d}{dC_{\mathrm{L,U}}}(y_{\mathrm{L,H}}) \end{pmatrix} = \begin{pmatrix} \chi_{\mathrm{H}}\beta_2(g_{\mathrm{H}} + mg_{\mathrm{L}}) & \chi_{\mathrm{L}}\beta_2(1-m)g_{\mathrm{H}} \\ \chi_{\mathrm{H}}\beta_2(1-m)g_{\mathrm{L}} & \chi_{\mathrm{L}}\beta_2(g_{\mathrm{L}} + mg_{\mathrm{H}}) \end{pmatrix}$$

$$= \frac{\chi\beta_2}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \left(\begin{pmatrix} f_{\mathrm{H}} r_{\mathrm{HL}}^2 & f_{\mathrm{H}} r_{\mathrm{HL}} \\ f_{\mathrm{L}} r_{\mathrm{HL}} & f_{\mathrm{L}} \end{pmatrix} + m r_{\mathrm{HL}} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \right)$$

$$T_{\mathrm{C,A}} = \begin{pmatrix} \frac{d}{dA_{\mathrm{H,U}}}(0) & \frac{d}{dA_{\mathrm{L,U}}}(0) \\ \frac{d}{dA_{\mathrm{H,U}}}(0) & \frac{d}{dA_{\mathrm{L,U}}}(0) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \mathbf{0}_{2,2}$$

$$T_{\mathrm{C,C}} = \begin{pmatrix} \frac{d}{dC_{\mathrm{H,U}}}(0) & \frac{d}{dC_{\mathrm{L,U}}}(0) \\ \frac{d}{dC_{\mathrm{H,U}}}(0) & \frac{d}{dC_{\mathrm{L,U}}}(0) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \mathbf{0}_{2,2}$$

To make further calculations slightly more convenient, let

$$T^* = \frac{1}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \left(\begin{pmatrix} f_{\mathrm{H}} r_{\mathrm{HL}}^2 & f_{\mathrm{H}} r_{\mathrm{HL}} \\ f_{\mathrm{L}} r_{\mathrm{HL}} & f_{\mathrm{L}} \end{pmatrix} + m r_{\mathrm{HL}} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \right)$$

Then $T_{\mathrm{A,A}} = \chi\beta_1 T^*$ and $T_{\mathrm{A,C}} = \chi\beta_2 T^*$.

Given these matrices, the basic reproduction number R_0 can be defined as the dominant eigenvalue of the Next-Generation Matrix with Large Domain[25]:

$$K_{\mathrm{L}} = -T\Sigma^{-1}. \tag{A.7}$$

There is a standard linear algebra result that says that given any 2x2 block matrix $\begin{pmatrix} W & X \\ Y & Z \end{pmatrix}$ for which Z, W, $(W - XZ^{-1}Y)$, and $(Z - YW^{-1}X)$ are all invertible, its inverse is:

$$\begin{pmatrix} W & X \\ Y & Z \end{pmatrix}^{-1} = \begin{pmatrix} (W - XZ^{-1}Y)^{-1} & -(W - XZ^{-1}Y)XZ^{-1} \\ -Z^{-1}Y(W - XZ^{-1}Y)^{-1} & (Z - YW^{-1}X)^{-1} \end{pmatrix}$$

Applying this result to Σ gives the equation:

$$\begin{aligned}
\Sigma^{-1} &= \begin{pmatrix} (\Sigma^{-1})_{\mathrm{A,A}} & (\Sigma^{-1})_{\mathrm{C,A}} \\ (\Sigma^{-1})_{\mathrm{C,C}} & (\Sigma^{-1})_{\mathrm{C,C}} \end{pmatrix} \\
&= \begin{pmatrix} \left(\Sigma_{\mathrm{A,A}} - \Sigma_{\mathrm{A,C}}\Sigma_{\mathrm{C,C}}^{-1}\Sigma_{\mathrm{C,A}}\right)^{-1} & -\left(\Sigma_{\mathrm{A,A}} - \Sigma_{\mathrm{A,C}}\Sigma_{\mathrm{C,C}}^{-1}\Sigma_{\mathrm{C,A}}\right)^{-1}\Sigma_{\mathrm{A,C}}\Sigma_{\mathrm{C,C}}^{-1} \\ -\Sigma_{\mathrm{C,C}}^{-1}\Sigma_{\mathrm{C,A}}\left(\Sigma_{\mathrm{A,A}} - \Sigma_{\mathrm{A,C}}\Sigma_{\mathrm{C,C}}^{-1}\Sigma_{\mathrm{C,A}}\right)^{-1} & \left(\Sigma_{\mathrm{C,C}} - \Sigma_{\mathrm{C,A}}\Sigma_{\mathrm{A,A}}^{-1}\Sigma_{\mathrm{A,C}}\right)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \left(\Sigma_{\mathrm{A,A}} - \mathbf{0}_{2,2}\Sigma_{\mathrm{C,C}}^{-1}\gamma_1\mathbf{I}_2\right)^{-1} & -\left(\Sigma_{\mathrm{A,A}} - \mathbf{0}_{2,2}\Sigma_{\mathrm{C,C}}^{-1}\gamma_1\mathbf{I}_2\right)^{-1}\mathbf{0}_{2,2}\Sigma_{\mathrm{C,C}}^{-1} \\ -\Sigma_{\mathrm{C,C}}^{-1}\gamma_1\mathbf{I}_2\left(\Sigma_{\mathrm{A,A}} - \mathbf{0}_{2,2}\Sigma_{\mathrm{C,C}}^{-1}\gamma_1\mathbf{I}_2\right)^{-1} & \left(\Sigma_{\mathrm{C,C}} - \gamma_1\mathbf{I}_2\Sigma_{\mathrm{A,A}}^{-1}\mathbf{0}_{2,2}\right)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_{\mathrm{A,A}}^{-1} & \mathbf{0}_{2,2} \\ -\gamma_1\Sigma_{\mathrm{C,C}}^{-1}\Sigma_{\mathrm{A,A}}^{-1} & \Sigma_{\mathrm{C,C}}^{-1} \end{pmatrix}
\end{aligned}$$

provided that $\Sigma_{\mathrm{A,A}}$ and $\Sigma_{\mathrm{C,C}}$ are both invertible.

Another standard linear algebra result is that for any 2x2 matrix:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = (ad-bc)^{-1} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

provided that $ac - bd \neq 0$. Therefore:

$$\begin{aligned}
\Sigma_{\mathrm{A,A}}^{-1} &= \begin{pmatrix} -(\nu_1 + f_{\mathrm{L}}\omega) & f_{\mathrm{H}}\omega \\ f_{\mathrm{L}}\omega & -(\nu_1 + f_{\mathrm{H}}\omega) \end{pmatrix}^{-1} \\
&= \frac{1}{(\nu_1 + f_{\mathrm{L}}\omega)(\nu_1 + f_{\mathrm{H}}\omega) - f_{\mathrm{H}}\omega f_{\mathrm{L}}\omega} \begin{pmatrix} -(\nu_1 + f_{\mathrm{H}}\omega) & -f_{\mathrm{H}}\omega \\ -f_{\mathrm{L}}\omega & -(\nu_1 + f_{\mathrm{L}}\omega) \end{pmatrix} \\
&= \frac{-1}{\nu_1(\nu_1+\omega)} \begin{pmatrix} (\nu_1 + f_{\mathrm{H}}\omega) & f_{\mathrm{H}}\omega \\ f_{\mathrm{L}}\omega & (\nu_1 + f_{\mathrm{L}}\omega) \end{pmatrix} \\
&= \frac{-1}{\nu_1} \begin{pmatrix} \frac{(\nu_1+f_{\mathrm{H}}\omega)}{(\nu_1+\omega)} & \frac{f_{\mathrm{H}}\omega}{(\nu_1+\omega)} \\ \frac{f_{\mathrm{L}}\omega}{(\nu_1+\omega)} & \frac{(\nu_1+f_{\mathrm{L}}\omega)}{(\nu_1+\omega)} \end{pmatrix} \\
&= \frac{-1}{\nu_1} \begin{pmatrix} f_{\mathrm{H}} + f_{\mathrm{L}}\psi_1 & f_{\mathrm{H}} - f_{\mathrm{H}}\psi_1 \\ f_{\mathrm{L}} - f_{\mathrm{L}}\psi_1 & f_{\mathrm{L}} + f_{\mathrm{H}}\psi_1 \end{pmatrix}
\end{aligned}$$

Similarly:

$$\begin{aligned}
\Sigma_{\mathrm{C,C}}^{-1} &= \frac{-1}{\nu_2(\nu_2+\omega)} \begin{pmatrix} (\nu_2 + f_{\mathrm{H}}\omega) & f_{\mathrm{H}}\omega \\ f_{\mathrm{L}}\omega & (\nu_2 + f_{\mathrm{L}}\omega) \end{pmatrix} \\
&= \frac{-1}{\nu_2} \begin{pmatrix} \frac{(\nu_2+f_{\mathrm{H}}\omega)}{(\nu_2+\omega)} & \frac{f_{\mathrm{H}}\omega}{(\nu_2+\omega)} \\ \frac{f_{\mathrm{L}}\omega}{(\nu_2+\omega)} & \frac{(\nu_2+f_{\mathrm{L}}\omega)}{(\nu_2+\omega)} \end{pmatrix} \\
&= \frac{-1}{\nu_2} \begin{pmatrix} f_{\mathrm{H}} + f_{\mathrm{L}}\frac{\psi_2}{\psi_1} & f_{\mathrm{H}} - f_{\mathrm{H}}\frac{\psi_2}{\psi_1} \\ f_{\mathrm{L}} - f_{\mathrm{L}}\frac{\psi_2}{\psi_1} & f_{\mathrm{L}} + f_{\mathrm{H}}\frac{\psi_2}{\psi_1} \end{pmatrix}
\end{aligned}$$

Therefore:

$$\begin{aligned}
\left(\Sigma^{-1}\right)_{\mathrm{C,A}} &= -\gamma_1 \Sigma_{\mathrm{A,A}}^{-1} \Sigma_{\mathrm{C,C}}^{-1} \\
&= -\gamma_1 \frac{-1}{\nu_1} \begin{pmatrix} f_{\mathrm{H}} + f_{\mathrm{L}}\psi_1 & f_{\mathrm{H}} - f_{\mathrm{H}}\psi_1 \\ f_{\mathrm{L}} - f_{\mathrm{L}}\psi_1 & f_{\mathrm{L}} + f_{\mathrm{H}}\psi_1 \end{pmatrix} \left(\frac{-1}{\nu_2}\right) \begin{pmatrix} f_{\mathrm{H}} + f_{\mathrm{L}}\frac{\psi_2}{\psi_1} & f_{\mathrm{H}} - f_{\mathrm{H}}\frac{\psi_2}{\psi_1} \\ f_{\mathrm{L}} - f_{\mathrm{L}}\frac{\psi_2}{\psi_1} & f_{\mathrm{L}} + f_{\mathrm{H}}\frac{\psi_2}{\psi_1} \end{pmatrix} \\
&= \frac{-\gamma_1}{\nu_1\nu_2} \begin{pmatrix} f_{\mathrm{H}} + f_{\mathrm{L}}\psi_2 & f_{\mathrm{H}} - f_{\mathrm{H}}\psi_2 \\ f_{\mathrm{L}} - f_{\mathrm{L}}\psi_2 & f_{\mathrm{L}} + f_{\mathrm{H}}\psi_2 \end{pmatrix}
\end{aligned}$$

Block matrices can also be used for matrix multiplication in the obvious way, and so:

$$\begin{aligned}
K_{\mathrm{L}} &= -T\Sigma^{-1} \\
&= \begin{pmatrix} \beta_1 T^* & \beta_2 T^* \\ 0_{2,2} & 0_{2,2} \end{pmatrix} \begin{pmatrix} -\Sigma_{\mathrm{A,A}}^{-1} & 0 \\ -\left(\Sigma^{-1}\right)_{\mathrm{C,A}} & -\Sigma_{\mathrm{C,C}}^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \chi\beta_1 T^* \Sigma_{\mathrm{A,A}}^{-1} + \chi\beta_2 T^* \left(-\left(\Sigma^{-1}\right)_{\mathrm{C,A}}\right) & \beta_2 T^* \Sigma_{\mathrm{C,C}}^{-1} \\ 0 & 0 \end{pmatrix}
\end{aligned}$$

Diekmann et al.[25] further note that the dominant eigenvalue of K_{L} is the same as the dominant eigenvalue of the Next-Generation Matrix K, which is the submatrix of K_{L} that contains only the rows and columns corresponding to states that individuals can have immediately upon infection (i.e., in this model, $A_{\mathrm{L,U}}$ and $A_{\mathrm{H,U}}$). That is to say:

$$
\begin{aligned}
K =&(K_{\mathrm{L}})_{\mathrm{A,A}} \\
=&\left(\chi\beta_1 T^* \Sigma^{-1}_{\mathrm{A,A}} + \chi\beta_2 T^* \left(-\left(\Sigma^{-1}\right)_{\mathrm{C,A}}\right)\right) \\
=&T^* \left(\chi\beta_1 \Sigma^{-1}_{\mathrm{A,A}} + \chi\beta_2 T^* \left(-\left(\Sigma^{-1}\right)_{\mathrm{C,A}}\right)\right) \\
=&T^* \left(\frac{\chi\beta_1}{\nu_1} \begin{pmatrix} f_{\mathrm{H}} + f_{\mathrm{L}}\psi_1 & f_{\mathrm{H}} - f_{\mathrm{H}}\psi_1 \\ f_{\mathrm{L}} - f_{\mathrm{L}}\psi_1 & f_{\mathrm{L}} + f_{\mathrm{H}}\psi_1 \end{pmatrix} + \frac{\beta_2\gamma_1}{\nu_1\nu_2} \begin{pmatrix} f_{\mathrm{H}} + f_{\mathrm{L}}\psi_2 & f_{\mathrm{H}} - f_{\mathrm{H}}\psi_2 \\ f_{\mathrm{L}} - f_{\mathrm{L}}\psi_2 & f_{\mathrm{L}} + f_{\mathrm{H}}\psi_2 \end{pmatrix} \right) \\
=&T^* \left((H_1 + H_2) \begin{pmatrix} f_{\mathrm{H}} & f_{\mathrm{H}} \\ f_{\mathrm{L}} & f_{\mathrm{L}} \end{pmatrix} + (H_1\psi_1 + H_2\psi_2) \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \right) \\
=&HT^* \left(\left(\begin{pmatrix} f_{\mathrm{H}} & f_{\mathrm{H}} \\ f_{\mathrm{L}} & f_{\mathrm{L}} \end{pmatrix} + \psi \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \right) \right) \\
=&H \left(\frac{1}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \left(\begin{pmatrix} f_{\mathrm{H}} r_{\mathrm{HL}}^2 & f_{\mathrm{H}} r_{\mathrm{HL}} \\ f_{\mathrm{L}} r_{\mathrm{HL}} & f_{\mathrm{L}} \end{pmatrix} + m r_{\mathrm{HL}} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \right) \right) \times \\
& \left(\begin{pmatrix} f_{\mathrm{H}} & f_{\mathrm{H}} \\ f_{\mathrm{L}} & f_{\mathrm{L}} \end{pmatrix} + \psi \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \right) \\
=&H(M_1 + \psi M_2 + m M_3 + \psi m M_4)
\end{aligned}
$$

where

$$
\begin{aligned}
M_1 =& \frac{1}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \begin{pmatrix} f_{\mathrm{H}} r_{\mathrm{HL}}^2 & f_{\mathrm{H}} r_{\mathrm{HL}} \\ f_{\mathrm{L}} r_{\mathrm{HL}} & f_{\mathrm{L}} \end{pmatrix} \begin{pmatrix} f_{\mathrm{H}} & f_{\mathrm{H}} \\ f_{\mathrm{L}} & f_{\mathrm{L}} \end{pmatrix} \\
=& \begin{pmatrix} \frac{f_{\mathrm{H}} r_{\mathrm{HL}}}{(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1))} & \frac{f_{\mathrm{H}} r_{\mathrm{HL}}}{(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1))} \\ \frac{f_{\mathrm{L}}}{(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1))} & \frac{f_{\mathrm{L}}}{(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1))} \end{pmatrix} \\
=& \begin{pmatrix} g_{\mathrm{H}} & g_{\mathrm{H}} \\ g_{\mathrm{L}} & g_{\mathrm{L}} \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
M_2 =& \frac{1}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \begin{pmatrix} f_{\mathrm{H}} r_{\mathrm{HL}}^2 & f_{\mathrm{H}} r_{\mathrm{HL}} \\ f_{\mathrm{L}} r_{\mathrm{HL}} & f_{\mathrm{L}} \end{pmatrix} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \\
=& \begin{pmatrix} \frac{f_{\mathrm{H}} f_{\mathrm{L}} r_{\mathrm{HL}}(r_{\mathrm{HL}}-1)}{(1+f_{\mathrm{H}}(r_{\mathrm{HL}}-1))} & \frac{-f_{\mathrm{H}}^2 r_{\mathrm{HL}}(r_{\mathrm{HL}}-1)}{(1+f_{\mathrm{H}}(r_{\mathrm{HL}}-1))} \\ \frac{f_{\mathrm{L}}^2 (r_{\mathrm{HL}}-1)}{(1+f_{\mathrm{H}}(r_{\mathrm{HL}}-1))} & \frac{-f_{\mathrm{H}} f_{\mathrm{L}}(r_{\mathrm{HL}}-1)}{(1+f_{\mathrm{H}}(r_{\mathrm{HL}}-1))} \end{pmatrix} \\
=& c_{\mathrm{V}}(\chi)^2 \begin{pmatrix} \frac{r_{\mathrm{HL}}}{r_{\mathrm{HL}}-1} & -\frac{f_{\mathrm{H}} r_{\mathrm{HL}}}{f_{\mathrm{L}}(r_{\mathrm{HL}}-1)} \\ \frac{f_{\mathrm{L}}}{f_{\mathrm{H}}(r_{\mathrm{HL}}-1)} & -\frac{1}{r_{\mathrm{HL}}-1} \end{pmatrix} \\
M_3 =& \frac{r_{\mathrm{HL}}}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \begin{pmatrix} f_{\mathrm{H}} & f_{\mathrm{H}} \\ f_{\mathrm{L}} & f_{\mathrm{L}} \end{pmatrix} \\
=& \mathbf{0}_{2,2} \\
M_4 =& \frac{r_{\mathrm{HL}}}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix} \\
=& \frac{r_{\mathrm{HL}}}{\left(1 + f_{\mathrm{H}}(r_{\mathrm{HL}} - 1)\right)^2} \begin{pmatrix} f_{\mathrm{L}} & -f_{\mathrm{H}} \\ -f_{\mathrm{L}} & f_{\mathrm{H}} \end{pmatrix}
\end{aligned}
$$

It's worth pausing here a moment to note one of the implications of this breakdown. The next-generation matrix can be expressed as the product of the homogeneous transmission potential (a scalar) and the sum of three matrix terms: One (M_1) that depends on neither the overall fraction of heterogeneity effect (ψ) nor the degree to which mixing is assortative (m); one (ψM_2) that is proportional to the overall fraction of heterogeneity effect, but does not depend on the degree to which mixing is assortative; and one ($\psi m M_4$) that is proportional to the product of the overall fraction of heterogeneity effect and the degree to which mixing is assortative. There is, however, no term that is proportional to the degree to which mixing is assortative, but does not depend on the overall fraction of heterogeneity effect: $mM_3 = 0$. This is because, if the overall fraction of heterogeneity effect is zero, then this means that the

population is effectively homogeneous, because the re-selection (ω) rate is infinite. Therefore, the extent to which mixing is assortative is irrelevant.

Adding these terms together, and introducing the scaled assortativity term $r = \frac{mr_{\mathrm{HL}}}{(1+f_{\mathrm{H}}(r_{\mathrm{HL}}-1))^2}$ from Table A.2 gives us:

$$K = H\begin{pmatrix} g_{\mathrm{H}}+\left(\left(\frac{r_{\mathrm{HL}}}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2+f_{\mathrm{L}}r\right)\psi & g_{\mathrm{H}}+\left(-\left(\frac{f_{\mathrm{H}}r_{\mathrm{HL}}}{f_{\mathrm{L}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2-f_{\mathrm{H}}r\right)\psi \\ g_{\mathrm{L}}+\left(\left(\frac{f_{\mathrm{L}}}{f_{\mathrm{H}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2-f_{\mathrm{L}}r\right)\psi & g_{\mathrm{L}}+\left(-\left(\frac{1}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2+f_{\mathrm{H}}r\right)\psi \end{pmatrix}$$

The dominant eigenvalue of a 2x2 matrix K is $\frac{\mathrm{tr}(K)+\sqrt{\mathrm{tr}(K)^2-4\mathrm{det}(K)}}{2}$, where $\mathrm{tr}(K) = K_{1,1} + K_{2,2}$ is the trace of K and $\det(K) = K_{1,1}K_{2,2} - K_{1,2}K_{2,1}$ is the determinant of K.

$$\begin{aligned}
\mathrm{tr}(K) =& H\left(g_{\mathrm{H}} + g_{\mathrm{L}} + \left(\left(\frac{r_{\mathrm{HL}}-1}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2 + (f_{\mathrm{L}}+f_{\mathrm{H}})r\right)\psi\right) \\
=& H\big(1 + \big(c_{\mathrm{V}}(\chi)^2 + r\big)\psi\big) \\
\det(K) =& H^2\left(g_{\mathrm{H}} + \left(\left(\frac{r_{\mathrm{HL}}}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2 + f_{\mathrm{L}}r\right)\psi\right)\times \\
& \left(g_{\mathrm{L}} + \left(-\left(\frac{1}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2 + f_{\mathrm{H}}r\right)\psi\right) \\
& - \left(g_{\mathrm{H}} + \left(-\left(\frac{f_{\mathrm{H}}r_{\mathrm{HL}}}{f_{\mathrm{L}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2 - f_{\mathrm{H}}r\right)\psi\right)\times \\
& \left(g_{\mathrm{L}} + \left(\left(\frac{f_{\mathrm{L}}}{f_{\mathrm{H}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2 - f_{\mathrm{L}}r\right)\psi\right) \\
=& H^2\psi r \\
\therefore R_0 =& \frac{\mathrm{tr}(K) + \sqrt{\mathrm{tr}(K)^2 - 4\mathrm{det}(K)}}{2} \\
=& H\left(\frac{1 + \big(c_{\mathrm{V}}(\chi)^2 + r\big)\psi + \sqrt{1 + 2\big(c_{\mathrm{V}}(\chi)^2 - r\big)\psi + \big(c_{\mathrm{V}}(\chi)^2 + r\big)^2\psi^2}}{2}\right)
\end{aligned}$$

Although this expression remains a bit opaque, three important points can be noted: First, it is monotonically increasing with respect to $c_{\mathrm{V}}(\chi)^2$ and r (defined in Table

A.2), representing measures of the relative degrees of risk heterogeneity and assortative mixing, respectively. Second, it is monotonically increasing with respect to ψ, and therefore monotonically decreasing with respect to ω. Finally, although three parameters (f_{H}, r_{HL}, and m) define the characteristics of behavioral heterogeneity apart from re-selection rate, R_0 depends on these through only two derived parameters, $c_{\mathrm{V}}(\chi)^2$ and r.

Once R_0 has been found, calculating the fraction of transmissions from EHI (ϕ) is relatively easy. Given K and its dominant eigenvalue (R_0), we can easily solve for the corresponding eigenvector v. If v is scaled to sum to 1, then its entries represent the fraction of newly infected individuals who are in each of the initial states ($A_{\mathrm{L,U}}$ and $A_{\mathrm{H,U}}$) in each generation during exponential growth. We can also decompose K into the sum of contributions to the NGM made by transmissions from EHI (K_1) and contributions made by transmissions from chronic infection (K_2), by simply omitting the proper terms from the algebra presented above. In the case of the model presented in this appendix, these matrices are:

$$K_1 = H_1 \begin{pmatrix} g_{\mathrm{H}}+\left(\left(\frac{r_{\mathrm{HL}}}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2+f_{\mathrm{L}}r\right)\psi_1 & g_{\mathrm{H}}+\left(-\left(\frac{f_{\mathrm{H}}r_{\mathrm{HL}}}{f_{\mathrm{L}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2-f_{\mathrm{H}}r\right)\psi_1 \\ g_{\mathrm{L}}+\left(\left(\frac{f_{\mathrm{L}}}{f_{\mathrm{H}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2-f_{\mathrm{L}}r\right)\psi_1 & g_{\mathrm{L}}+\left(-\left(\frac{1}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2+f_{\mathrm{H}}r\right)\psi_1 \end{pmatrix}$$

$$K_2 = H_2 \begin{pmatrix} g_{\mathrm{H}}+\left(\left(\frac{r_{\mathrm{HL}}}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2+f_{\mathrm{L}}r\right)\psi_2 & g_{\mathrm{H}}+\left(-\left(\frac{f_{\mathrm{H}}r_{\mathrm{HL}}}{f_{\mathrm{L}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2-f_{\mathrm{H}}r\right)\psi_2 \\ g_{\mathrm{L}}+\left(\left(\frac{f_{\mathrm{L}}}{f_{\mathrm{H}}(r_{\mathrm{HL}}-1)}\right)c_{\mathrm{V}}(\chi)^2-f_{\mathrm{L}}r\right)\psi_2 & g_{\mathrm{L}}+\left(-\left(\frac{1}{r_{\mathrm{HL}}-1}\right)c_{\mathrm{V}}(\chi)^2+f_{\mathrm{H}}r\right)\psi_2 \end{pmatrix}$$

The total number of transmissions per unit of infected population in one generation during exponential growth will then be $||K_1 v||_1$ (the sum of the two entries of the column vector $K_1 v$) from EHI, and $\|K_2 v\|_1$ from chronic infection. Consequently the fraction of transmissions from EHI is:

$$\phi = \frac{\|K_1 v\|_1}{\|K_1 v\|_1 + \|K_2 v\|_1} = \frac{\|K_1 v\|_1}{\|K v\|_1} \tag{A.8}$$

This can be expressed algebraically in terms of the original parameters, but the resulting equation is neither brief nor enlightening.

The formula for R_0 is also useful in calculating the treatment rate required to achieve elimination of ongoing transmission (τ_E): Because elimination will occur precisely when $R_0 = 1$, τ_E can be found by solving for the value of τ that makes $R_0 = 1$. Although it is technically possible to do this algebraically, it is simpler to do it numerically, and that is what we do. All code used can be obtained by the means discussed in Appendix B.

A.4.7 Construction of Figure 3.4

As noted in Table 3.3, we parameterize the per-act transmissibility during EHI and chronic infection by β_2 and ζ, or, more descriptively, the per-act transmissibility during chronic infection and the relative transmissibility during EHI (i.e. the ratio of β_1 to β_2). With this parameterization, ϕ depends on ζ, but not on β_2, while the endemic prevalence depends on both. Consequently, we use the parameters of Table A.1 (except for ω and ζ), and for each value of ω, we first find a value for ζ that will result in a ϕ of 0.447. Then, using that value for ζ, we find a value for β_2 that will result in an endemic prevalence of 0.2. With a full parameter set, we then find R_0 and τ_E using the methods detailed above.

A.5 Supplementary Results

A.5.1 Relationship between the basic reproduction number (R_0), the fraction of transmission from early infection (ϕ), and the effective treatment rate required for elimination of ongoing transmission (τ_E)

If we denote the basic reproduction number in the presence of diagnosis and treatment as $R_0(\tau)$, then elimination will occur precisely when $R_0(\tau)$ is reduced to

1, i.e.

$$R_0(\tau_E) = 1. \tag{A.9}$$

Our previous equation for R_0 easily extends to:

$$R_0 = H(\tau)\left(\frac{1+\left(c_{\mathrm{V}}(\chi)^2+r\right)\psi(\tau)+\sqrt{1+2\left(c_{\mathrm{V}}(\chi)^2-r\right)\psi(\tau)+\left(c_{\mathrm{V}}(\chi)^2+r\right)^2\psi(\tau)^2}}{2}\right) \tag{A.10}$$

In submodels without episodic risk ($\omega = 0$), $\psi = 1$. Consequently, the above equation simplifies to:

$$R_0 = H(\tau)\left(\frac{1+\left(c_{\mathrm{V}}(\chi)^2+r\right)+\sqrt{1+2\left(c_{\mathrm{V}}(\chi)^2-r\right)+\left(c_{\mathrm{V}}(\chi)^2+r\right)^2}}{2}\right) \tag{A.11}$$

From this equation and the equations given in Table A.2, it follows that:

$$R_0(\tau) = R_0(0)\left(\phi\left(\frac{\mu+\gamma_1}{\mu+\gamma_1+\tau}\right)+(1-\phi)\left(\frac{\mu+\gamma_1}{\mu+\gamma_1+\tau}\right)\left(\frac{\mu+\gamma_2}{\mu+\gamma_2+\tau}\right)\right) \tag{A.12}$$

By solving this equation for $R_0(\tau_E) = 1$, we obtain an equation for the effective treatment rate required to achieve elimination (τ_E) that depends on ζ and the behavioral parameters (f_{H}, r_{HL}, and m) only through their effects on $R_0(0)$ (henceforth simply R_0) and the fraction of transmissions from early infection (ϕ):

$$\tau_{\mathrm{E}} = \frac{R_0\phi(\mu+\gamma_1)-(2\mu+\gamma_1+\gamma_2)+\sqrt{(R_0\phi(\mu+\gamma_1)-(2\mu+\gamma_1+\gamma_2))^2+4(R_0-1)(\mu+\gamma_1)(\mu+\gamma_2)}}{2} \tag{A.13}$$

This result (depicted graphically in Figure A.1) shows that τ_E depends strongly on both R_0 and ϕ (being monotonically increasing with respect to both). However, the effects of R_0 are somewhat stronger than those of ϕ – if $\phi = 0$, and $R_0 > 1$, then $\tau_E > 0$, but if $\phi > 0$, and $R_0 = 1$, then $\tau_E = 0$.

As we discuss further below, this algebraic result does not hold in the full model, or in any submodel that includes episodic risk ($\omega \neq 0$). In the presence of episodic risk, behavioral parameters can have effects on τ_E even when R_0 and ϕ are both fixed. Nevertheless, the broader qualitative result, that τ_E depends strongly on both R_0 and ϕ, but somewhat more strongly on R_0 than on ϕ, continues to apply.

A.5.2 Effects of re-selection rate (ω) on fraction of transmissions from EHI (ϕ)

In the absence of assortative mixing, and during exponential growth, the average number of secondary transmissions from EHI per index case is $(1 + c_V(\chi)^2\psi_1)H_1$, and the average number of secondary transmission during chronic infection per index case is $(1 + c_V(\chi)^2\psi_2)H_2$. The fraction of transmissions (made by all individuals in a generation) that are from EHI is therefore:

$$\phi = \frac{(1 + c_V(\chi)^2\psi_1)H_1}{(1 + c_V(\chi)^2\psi_1)H_1 + (1 + c_V(\chi)^2\psi_2)H_2} \tag{A.14}$$

For conceptual understanding, we are less interested in the exact value of ϕ at a particular ω (although this can be calculated from the equation above in a straightforward manner) than we are in the conditions under which ϕ is increasing or decreasing with respect to ω, i.e. we are interested in the sign of $\frac{d\phi}{d\omega}$. This is easily shown to be the same as the sign of $\frac{d}{d\omega}(\frac{(1+c_V(\chi)^2\psi_1)H_1}{(1+c_V(\chi)^2\psi_2)H_2})$. From this, and the equations given in Table A.2, it is a matter of straightforward algebra to show that $\frac{d}{d\omega}(\frac{(1+c_V(\chi)^2\psi_1)H_1}{(1+c_V(\chi)^2\psi_2)H_2})$ is always decreasing, with a zero at:

$$\omega = \sqrt{\nu_1\nu_2(1 + c_V(\chi)^2)}$$

Thus, the fraction of transmissions from EHI is first increasing with respect to the re-selection rate, then decreasing, as noted in the Chapter III. The maximum is attained when the re-selection rate is the product of the square root of the multiplicative boost to R_0 due to static risk heterogeneity and the harmonic mean of the net rates of leaving (1) the untreated and acutely infected state and (2) the untreated and chronically infected state.

A.5.3 Conversion of effective treatment rate to "actual" diagnosis rate

For a highly simplified calculation relating our concept of a single "effective treatment rate" to actual diagnosis rates, we will make the following simplifying assumptions:

1. Time from infection to diagnosis follows an exponential distribution with rate parameter δ, the diagnosis rate.
2. Time from diagnosis to entry into treatment follows an exponential distribution with rate parameter ϵ.
3. Time from entry into treatment to viral suppression follows an exponential distribution with rate parameter σ.
4. There is no virological failure after viral suppression.
5. The "effective treatment rate" τ for a treatment cascade as defined above is the rate parameter for an exponential distribution having the same mean as the actual distribution of times from infection to viral suppression, i.e.

$$\tau = \frac{1}{\frac{1}{\delta} + \frac{1}{\epsilon} + \frac{1}{\sigma}}$$
$$\therefore \delta = \frac{1}{\frac{1}{\tau} - \frac{1}{\epsilon} - \frac{1}{\sigma}}$$

None of these assumptions are actually true. Consequently, this is a poor model to use if the goal is to convert a specific effective treatment rate to a specific diagnostic rate. Nevertheless, it is useful for getting a qualitative sense of what the difference between two effective treatment rates required for elimination implies about the difference between the diagnostic rates required for elimination in their respective models. With that in mind, we proceed. A recent paper estimated the rate of initiating treatment following diagnosis as 77.2% within 3 months[124]. Combining

this estimate with the above assumptions gives:

$$\epsilon = -4 \ln 0.228 = 5.91$$

Crawford et al.[23] report that the median time from initiating treatment to viral suppression for all patients was 1.03 years, and that for patients with optimal retention in care, it was 0.87 years. These figures give us (respectively):

$$\sigma_{\text{all}} = \frac{-ln0.5}{1.03} = 0.673$$
$$\sigma_{\text{optimal}} = \frac{-ln0.5}{0.87} = 0.797$$

Using the figure for all patients, $\tau_E = 0.0635$ implies that $\delta_E = 0.0710$. In contrast, $\tau_E = 0.891$ implies that the mean time from infection to viral suppression must be only 1.12 years. But even if we assume that perfect retention in care could be achieved for all patients, $\sigma \approx 0.797$ implies that the mean time from entry into care to viral suppression is approximately 1.25 years, meaning that even instantaneous diagnosis and entry into care upon infection would not be sufficient to eliminate transmission through UT&T alone.

A.5.4 Verification of qualitative results using an elaborated model of disease progression

In order to verify the robustness of the model to the relaxation of one of our simplifying assumptions, we tested a modified version in which we replaced our two-stage simplification of the natural history of HIV infection with a more realistic version of that natural history, based on the results of a study of HIV transmission in Lilongwe, Malawi[100]. In this version of the model, HIV progression consists of 9 stages, each of which has its own average duration, (relative) per-contact transmission rate, and identification as an acute stage or a chronic stage. The model used for the analysis of the Lilongwe data was Bayesian in nature; for our purposes, we used the

posterior modes for each of the relevant parameters. We omitted the last of their 10 stages, corresponding to late AIDS, because they modeled this stage as having a transmission rate of 0, due to the cessation of sexual activity as the result of AIDS symptoms. Because our model does not include long-term partnerships, and only considers the sexually active population, individuals who have permanently ceased sexual activity are irrelevant to it. For consistency with our main model, we coded the progression rate for effectively treated individuals as being equal to the untreated rate for acute stages, and 0 for chronic stages. The accuracy of this approach is, for the purposes of Chapter III and this appendix, irrelevant, as all of our analyses involve either the absence of effective treatment (when calculating the endemic prevalence prior to intervention) or conditions during exponential growth (when calculating the basic reproduction number or fraction of transmissions from EHI). Consequently, the fate of effectively treated individuals has no impact on any of our results. The initial parameters for each stage are given in Table A.3. In order to perform analyses analogous to those we performed on our primary model, we calculated an average per-contact transmission rate for each of acute and chronic infection, by weighting the transmission rate for each stage by the expected time spent in that stage. We then treated these averages analogously to β_1 and β_2 in the main model, respectively. So, for example, where we allowed β_1 to vary in the main model, we allowed the transmission rates for all acute stages to vary, but required that the ratios between those transmission rates remain constant.

Figures corresponding to Figures 3.1, 3.2, 3.3, 3.4, and A.1, with no changes apart from the replacement of the main model with the elaborated model, are presented as Figures A.2, A.3, A.4, A.5, and A.6, respectively. It is readily seen that, although there are quantitative differences, these figures are all qualitatively the same as their

main-model counterparts. This shows that our qualitative conclusions are robust to the relaxation of our simplifying assumption of a two-stage natural history of infection. Similarly, the only equations that we present in Chapter III or in this appendix that do not still hold for the elaborated model are those that explicitly indicate only two stages.

APPENDIX B

Obtaining Source Code Used for Simulations and Calculations

All source code used in simulations and calculations whose results are presented in this dissertation is available to download from a git repository located at:

https://github.com/cjh57/dissertation

Should any errors in the text of this dissertation be discovered after publication, a list of errata will be maintained there as well. Source code can also be obtained, and questions answered, by sending an email to the author at chrishen@umich.edu.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Frithjofna Abbink, Anne M Buisman, Gerda Doornbos, Jan Woldman, Tjeerd G Kimman, and Marina AE Conyn-van Spaendonck. Poliovirus-specific memory immunity in seronegative elderly people does not protect against virus excretion. *Journal of Infectious Diseases*, 191(6):990–999, 2005.

[2] Laith J Abu-Raddad and Ira M Longini Jr. No HIV stage is dominant in driving the HIV epidemic in sub-Saharan Africa. *AIDS*, 22(9):1055–1061, 2008.

[3] Onoja Matthew Akpa and Benjamin Agboola Oyejola. Modeling the transmission dynamics of HIV/AIDS epidemics: an introduction and a review. *The Journal of Infection in Developing Countries*, 4(10):597–608, 2010.

[4] Muhammad Masroor Alam, Salmaan Sharif, Shahzad Shaukat, Mehar Angez, Adnan Khurshid, Lubna Rehman, and Syed Sohail Zahoor Zaidi. Genomic surveillance elucidates persistent wild poliovirus transmission during 2013-2015 in major reservoir areas of Pakistan. *Clinical Infectious Diseases*, page civ831, 2015.

[5] Shah Jamal Alam, Ethan Romero-Severson, Jong-Hoon Kim, Gilbert Emond, and James S Koopman. Dynamic sex roles among men who have sex with men and transmissions from primary HIV infection. *Epidemiology (Cambridge, Mass.)*, 21(5):669, 2010.

[6] RM Anderson, GF Medley, RM May, and AM Johnson. A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *Mathematical Medicine and Biology: a Journal of the IMA*, 3(4):229–263, 1986.

[7] Roy M Anderson and Robert M May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[8] E Anis, E Kopel, S R Singer, E Kaliner, L Moerman, J Moran-Gilad, D Sofer, Y Manor, L M Shulman, E Mendelson, M Gdalevich, B Lev, R Gamzu, and I Grotto. Insidious reintroduction of wild poliovirus into Israel, 2013 Detection of wild poliovirus in Israel. *Euro Surveillance*, 18(38), 2013.

[9] Sophie Arie. Polio virus spreads from Syria to Iraq. *BMJ: British Medical Journal*, 348, 2014.

[10] Humayun Asghar, Ousmane M Diop, Goitom Weldegebriel, Farzana Malik, Sushmitha Shetty, Laila El Bassioni, Adefunke O Akande, Eman Al Maamoun, Sohail Zaidi, Adekunle J Adeniji, et al. Environmental surveillance for polioviruses in the Global Polio Eradication Initiative. *Journal of Infectious Diseases*, 210(suppl 1):S294–S303, 2014.

[11] R Bruce Aylward and Ala Alwan. Polio in Syria. *The Lancet*, 383(9916):489–491, 2014.

[12] Matthew R Behrend, Hao Hu, Karima R Nigmatulina, and Philip Eckhoff. A quantitative survey of the literature on poliovirus infection and immunity. *International Journal of Infectious Diseases*, 18:4–13, 2014.

[13] Daniela Bezemer, Frank de Wolf, Maarten C Boerlijst, Ard van Sighem, T Deirdre Hollingsworth, and Christophe Fraser. 27 years of the HIV epidemic amongst men having sex with men in the Netherlands: an in depth mathematical model-based analysis. *Epidemics*, 2(2):66–79, 2010.

[14] Sally M Blower, Gotfried JP van Griensven, and Edward H Kaplan. An analysis of the process of human immunodeficiency virus sexual risk behavior change. *Epidemiology*, pages 238–242, 1995.

[15] SM Blower, MC Samuel, and JA Wiley. Sex, power laws, and HIV transmission. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 5(6):633–634, 1992.

[16] George EP Box, Norman Richard Draper, et al. *Empirical model-building and response surfaces*. Wiley New York, 1987.

[17] Bluma Brenner and Erica EM Moodie. HIV sexual networks: the Montreal experience. *Statistical Communications in Infectious Diseases*, 4(1), 2012.

[18] Bluma G Brenner, Michel Roger, Jean-Pierre Routy, Daniela Moisi, Michel Ntemgwa, Claudine Matte, Jean-Guy Baril, Réjean Thomas, Danielle Rouleau, Julie Bruneau, et al. High rates of forward transmission events after acute/early HIV-1 infection. *The Journal of infectious diseases*, 195(7):951–959, 2007.

[19] Bluma G Brenner, Michel Roger, David Stephens, Daniela Moisi, Isabelle Hardy, Jonathan Weinberg, Reuven Turgel, Hugues Charest, James Koopman, Mark A Wainberg, et al. Transmission clustering drives the onward spread of the HIV epidemic among men who have sex with men in Quebec. *Journal of Infectious Diseases*, 204(7):1115–1119, 2011.

[20] Cara C Burns, Ousmane M Diop, Roland W Sutter, and Olen M Kew. Vaccine-derived polioviruses. *Journal of Infectious Diseases*, 210(suppl 1):S283–S293, 2014.

[21] Myron S Cohen, Christopher Dye, Christophe Fraser, William C Miller, Kimberly A Powers, and Brian G Williams. HIV treatment as prevention: debate and commentary–will early infection compromise treatment-as-prevention strategies? *PLoS medicine*, 9(7):e1001232, 2012.

[22] Myron S Cohen, George M Shaw, Andrew J McMichael, and Barton F Haynes. Acute HIV-1 infection. *New England Journal of Medicine*, 364(20):1943–1954, 2011.

[23] Timothy N Crawford, Wayne T Sanderson, and Alice Thornton. Impact of poor retention in HIV medical care on time to viral load suppression. *Journal of the International Association of Providers of AIDS Care (JIAPAC)*, 13(3):242–249, 2014.

[24] Sara M Debanne and Douglas Y Rowland. Statistical certification of eradication of poliomyelitis in the Americas. *Mathematical biosciences*, 150(1):83–103, 1998.

[25] O Diekmann, JAP Heesterbeek, and MG Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, page rsif20090386, 2009.

[26] Klaus Dietz. The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research*, 2(1):23–41, 1993.

[27] Joseph DiStefano III. *Dynamic systems biology modeling and simulation*. Academic Press, 2015.

[28] Radboud J Duintjer Tebbens, Dominika A Kalkowska, Steven GF Wassilak, Mark A Pallansch, Stephen L Cochi, and Kimberly M Thompson. The potential impact of expanding target age groups for polio immunization campaigns. *BMC infectious diseases*, 14(1):45, 2014.

[29] Radboud J Duintjer Tebbens, Mark A Pallansch, Konstantin M Chumakov, Neal A Halsey, Tapani Hovi, Philip D Minor, John F Modlin, Peter A Patriarca, Roland W Sutter, Peter F Wright, et al. Expert review on poliovirus immunity and transmission. *Risk Analysis*, 33(4):544–605, 2013.

[30] Radboud J Duintjer Tebbens, Mark A Pallansch, Konstantin M Chumakov, Neal A Halsey, Tapani Hovi, Philip D Minor, John F Modlin, Peter A Patriarca, Roland W Sutter, Peter F Wright, et al. Review and assessment of poliovirus immunity and transmission: Synthesis of knowledge gaps and identification of research needs. *Risk Analysis*, 33(4):606–646, 2013.

[31] Radboud J Duintjer Tebbens, Mark A Pallansch, Dominika A Kalkowska, Steven GF Wassilak, Stephen L Cochi, and Kimberly M Thompson. Characterizing poliovirus transmission and evolution: Insights from modeling experiences with wild and vaccine-related polioviruses. *Risk Analysis*, 33(4):703–749, 2013.

[32] Radboud J Duintjer Tebbens, Mark A Pallansch, Jong-Hoon Kim, Cara C Burns, Olen M Kew, M Steven Oberste, Ousmane M Diop, Steven GF Wassilak, Stephen L Cochi, and Kimberly M Thompson. Oral poliovirus vaccine evolution and insights relevant to modeling the risks of circulating vaccine-derived polioviruses (cVDPVs). *Risk analysis*, 33(4):680–702, 2013.

[33] Radboud J Duintjer Tebbens, Mark A Pallansch, Steven GF Wassilak, Stephen L Cochi, and Kimberly M Thompson. Combinations of quality and frequency of immunization activities to stop and prevent poliovirus transmission in the high-risk area of northwest Nigeria. *PLoS One*, 10(6):e0130123, 2015.

[34] Radboud J Duintjer Tebbens and Kimberly M Thompson. Modeling the potential role of inactivated poliovirus vaccine to manage the risks of oral poliovirus vaccine cessation. *Journal of Infectious Diseases*, 210(suppl 1):S485–S497, 2014.

[35] Radboud J Duintjer Tebbens and Kimberly M Thompson. Managing the risk of circulating vaccine-derived poliovirus during the endgame: Oral poliovirus vaccine needs. *BMC infectious diseases*, 15(1):390, 2015.

[36] Jeffrey W Eaton and Timothy B Hallett. Why the proportion of transmission during early-stage HIV infection does not predict the long-term impact of treatment on HIV incidence. *Proceedings of the national academy of sciences*, 111(45):16202–16207, 2014.

[37] Jeffrey W Eaton, Timothy B Hallett, and Geoffrey P Garnett. Concurrent sexual partnerships and primary HIV infection: a critical interaction. *AIDS and Behavior*, 15(4):687–692, 2011.

[38] M Eichner and K Dietz. Eradication of poliomyelitis: when can one be sure that polio virus transmission has been terminated? *American journal of epidemiology*, 143(8):816–822, 1996.

[39] Joseph NS Eisenberg, James Trostle, Reed JD Sorensen, and Katherine F Shields. Toward a systems approach to enteric pathogen transmission: from individual independence to community interdependence. *Annual review of public health*, 33:239–257, 2012.

[40] Andrew Etsano. Environmental isolation of circulating vaccine-derived poliovirus after interruption of wild poliovirus transmission–Nigeria, 2016. *MMWR. Morbidity and Mortality Weekly Report*, 65, 2016.

[41] Michael Famulare. Has wild poliovirus been eliminated from Nigeria? *PloS one*, 10(8):e0135765, 2015.

[42] Michael Famulare, Guillaume Chabot-Couture, Philip A Eckhoff, Hil Lyons, Kevin A McCarthy, and Christian Selinger. How polio vaccination affects poliovirus transmission. *bioRxiv*, page 084012, 2016.

[43] Paul EM Fine and Ilona AM Carneiro. Transmissibility and persistence of oral polio vaccine viruses: implications for the Global Poliomyelitis Eradication Initiative. *American journal of epidemiology*, 150(10):1001–1021, 1999.

[44] Warren A Furumoto and Ray Mickey. A mathematical model for the infectivity-dilution curve of tobacco mosaic virus: theoretical considerations. *Virology*, 32(2):216–223, 1967.

[45] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[46] Global Polio Eradication Initiative. Guidelines on environmental surveillance for detection of polioviruses. Technical Report March, 2015.

[47] Global Polio Eradication Initiative. Polio Environmental Surveillance Expansion Plan. Technical Report April, 2015.

[48] Global Polio Eradication Initiative. Surveillance Indicators: Acute flaccid Paralysis (AFP) Surveillance, Polio Today, 2017.

[49] M Gabriela M Gomes, Lisa J White, and Graham F Medley. The reinfection threshold. *Journal of theoretical biology*, 236(1):111–113, 2005.

[50] Steven M Goodreau and Matthew R Golden. Biological and demographic causes of high HIV and sexually transmitted disease prevalence in men who have sex with men. *Sexually transmitted infections*, 83(6):458–462, 2007.

[51] Nicholas C Grassly. The final stages of the global eradication of poliomyelitis. *Phil. Trans. R. Soc. B*, 368(1623):20120140, 2013.

[52] Nicholas C Grassly, Hamid Jafari, Sunil Bahl, Raman Sethi, Jagadish M Deshpande, Chris Wolff, Roland W Sutter, and R Bruce Aylward. Waning intestinal immunity after vaccination with oral poliovirus vaccines in India. *Journal of Infectious Diseases*, page jis241, 2012.

[53] Sunetra Gupta, Roy M Anderson, and Robert M May. Networks of sexual contacts: Implication for the pattern of spread. *AIDS*, 3(12), 1989.

[54] Charles N Haas, Joan B Rose, and Charles P Gerba. *Quantitative microbial risk assessment*. John Wiley & Sons, 1999.

[55] Jose E Hagan, SG Wassilak, Allen S Craig, Rudolf H Tangermann, Ousmane M Diop, Cara C Burns, and Arshad Quddus. Progress toward polio eradication—worldwide, 2014–2015. *MMWR. Morbidity and mortality weekly report*, 64(19):527–531, 2015.

[56] H Irene Hall, Ruiguang Song, Philip Rhodes, Joseph Prejean, Qian An, Lisa M Lee, John Karon, Ron Brookmeyer, Edward H Kaplan, Matthew T McKenna, et al. Estimation of HIV incidence in the United States. *JAMA*, 300(5):520–529, 2008.

[57] H Irene Hall, Ruiguang Song, Célia Landmann Szwarcwald, and Timothy Green. Brief report: Time from infection with the human immunodeficiency virus to diagnosis, United States. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 69(2):248–251, 2015.

[58] Christopher J Henry and James S Koopman. Strong influence of behavioral dynamics on the ability of testing and treating HIV to stop transmission. *Scientific reports*, 5, 2015.

[59] Herbert W Hethcote and James W Van Ark. Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. *Mathematical Biosciences*, 84(1):85–118, 1987.

[60] DL Heymann, EM De Gourville, and RB Aylward. Protecting investments in polio eradication: the past, present and future of surveillance for acute flaccid paralysis. *Epidemiology and infection*, 132(05):779–780, 2004.

[61] T Déirdre Hollingsworth, Roy M Anderson, and Christophe Fraser. HIV-1 transmission, by stage of infection. *The Journal of infectious diseases*, 198(5):687–693, 2008.

[62] T Hovi, LM Shulman, H Van Der Avoort, J Deshpande, M Roivainen, EM De Gourville, et al. Role of environmental poliovirus surveillance in global polio eradication and beyond. *Epidemiology and infection*, 140(1):1, 2012.

[63] Gareth J Hughes, Esther Fearnhill, David Dunn, Samantha J Lycett, Andrew Rambaut, Andrew J Leigh Brown, UK HIV Drug Resistance Collaboration, et al. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS pathogens*, 5(9):e1000590, 2009.

[64] Global Polio Eradication Initiative. Polio eradication and endgame midterm review, July 2015. *Geneva, Switzerland: World Health Organization*, 2015.

[65] John A Jacquez, James S Koopman, Carl P Simon, and Ira M Longini Jr. Role of the primary infection in epidemics of HIV infection in gay cohorts. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 7(11):1169–1184, 1994.

[66] John A Jacquez, Carl P Simon, James Koopman, Lisa Sattenspiel, and Timothy Perry. Modeling and analyzing HIV transmission: the effect of contact patterns. *Mathematical Biosciences*, 92(2):119–199, 1988.

[67] Hamid Jafari, Jagadish M Deshpande, Roland W Sutter, Sunil Bahl, Harish Verma, Mohammad Ahmad, Abhishek Kunwar, Rakesh Vishwakarma, Ashutosh Agarwal, Shilpi Jain, et al. Efficacy of inactivated poliovirus vaccine in India. *Science*, 345(6199):922–925, 2014.

[68] Ticha Johnson Muluh, Abdullahi Walla Hamisu, Kehinde Craig, Pascal Mkanda, Etsano Andrew, Johnson Adeniji, Adefunke Akande, Audu Musa, Isiaka Ayodeji, Gumede Nicksy, et al. Contribution of environmental surveillance toward interruption of poliovirus transmission in Nigeria, 2012–2015. *Journal of Infectious Diseases*, page jiv767, 2016.

[69] Dominika A Kalkowska, Radboud J Duintjer Tebbens, Mark A Pallansch, Stephen L Cochi, Steven GF Wassilak, and Kimberly M Thompson. Modeling undetected live poliovirus circulation after apparent interruption of transmission: Implications for surveillance and vaccination. *BMC infectious diseases*, 15(1):66, 2015.

[70] Dominika A Kalkowska, Radboud J Duintjer Tebbens, and Kimberly M Thompson. The probability of undetected wild poliovirus circulation after apparent global interruption of transmission. *American journal of epidemiology*, 175(9):936–949, 2012.

[71] William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 115, pages 700–721. The Royal Society, 1927.

[72] Jong-Hoon Kim, Rick L Riolo, and James S Koopman. HIV transmission by stage of infection and pattern of sexual partnerships. *Epidemiology (Cambridge, Mass.)*, 21(5):676, 2010.

[73] A.A. King, E. Ionides, C. Breto, S.P. Ellner, M.J. Ferrari, Kendall, B.E., M. Lavine, D. Nguyen, D.C. Reuman, H. Wearing, S.N. Wood, and S. Funk. R software package "pomp", 2016.

[74] J Koopman and B Foxman. Transmission modeling to enhance surveillance system function. *Transforming Public Health Surveillance. Elsevier*, 2016.

[75] J Koopman, P Singh, and E Ionides. Transmission modeling to enhance surveillance system function. *Transforming Public Health Surveillance. Elsevier*, 2014.

[76] James S Koopman et al. Infection transmission science and models. *Japanese journal of infectious diseases*, 58(6):S, 2005.

[77] James S Koopman, Christopher J Henry, JH Park, MC Eisenberg, EL Ionides, and JN Eisenberg. Dynamics affecting the risk of silent circulation when oral polio vaccination is stopped. *Epidemics*, 2017.

[78] James S Koopman, John A Jacquez, Gavin W Welch, Carl P Simon, Betsy Foxman, Stephen M Pollock, Daniel Barth-Jones, Andrew L Adams, Kenneth Lange, et al. The role of early HIV infection in the spread of HIV through populations. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 14(3):249–258, 1997.

[79] James S Koopman, Carl P Simon, and Chris P Riolo. When to control endemic infections by focusing on high-risk groups. *Epidemiology*, 16(5):621–627, 2005.

[80] H Kraan, P Soema, JP Amorij, and G Kersten. Intranasal and sublingual delivery of inactivated polio vaccine. *Vaccine*, 35(20):2647, 2017.

[81] Mirjam E Kretzschmar, Maarten F Schim van der Loeff, Paul J Birrell, Daniela De Angelis, and Roel A Coutinho. Prospects of elimination of HIV with test-and-treat strategy. *Proceedings of the National Academy of Sciences*, 110(39):15538–15543, 2013.

[82] Fraser Lewis, Gareth J Hughes, Andrew Rambaut, Anton Pozniak, and Andrew J Leigh Brown. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS medicine*, 5(3):e50, 2008.

[83] EF Maes, OM Diop, J Jorba, S Chavan, RH Tangermann, and SG Wassilak. Surveillance systems to track progress toward polio eradication—worldwide, 2015-2016. *MMWR. Morbidity and mortality weekly report*, 66(13):359, 2017.

[84] Y Manor, LM Shulman, E Kaliner, M Hindiyeh, D Ram, D Sofer, J Moran-Gilad, B Lev, I Grotto, R Gamzu, et al. Intensified environmental surveillance supporting the response to wild poliovirus type 1 silent circulation in Israel, 2013. *Euro Surveill*, 19(7):20708, 2014.

[85] RM May and RM Anderson. Transmission dynamics of HIV infection. *Nature*, 326(6109):137, 1987.

[86] Bryan T Mayer, Joseph NS Eisenberg, Christopher J Henry, M Gabriela M Gomes, Edward L Ionides, and James S Koopman. Successes and shortcomings of polio eradication: a transmission modeling analysis. *American journal of epidemiology*, 177(11):1236–1245, 2013.

[87] Michelle Morales. Progress toward polio eradication—worldwide, 2015–2016. *MMWR. Morbidity and mortality weekly report*, 65, 2016.

[88] Annett Nold. Heterogeneity in disease-transmission modeling. *Mathematical Biosciences*, 52(3-4):227–240, 1980.

[89] Brian A Nosek, George Alter, George C Banks, D Borsboom, SD Bowman, SJ Breckler, S Buck, Christopher D Chambers, G Chin, G Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

[90] World Health Organization et al. Polio eradication and endgame strategic plan 2013–2018. *Global Polio Eradication Initiative, working draft*, 23:1–99, 2013.

[91] World Health Organization et al. Who global action plan to minimize polio virus facility-associated risk (GAPIII). *France: WHO*, pages 8–16, 2015.

[92] Miguel O'Ryan, Ananda S Bandyopadhyay, Rodolfo Villena, Mónica Espinoza, José Novoa, William C Weldon, M Steven Oberste, Steve Self, Bhavesh R Borate, Edwin J Asturias, et al. Inactivated poliovirus vaccine given alone or in a sequential schedule with bivalent oral poliovirus vaccine in Chilean infants: a randomised, controlled, open-label, phase 4, non-inferiority study. *The Lancet Infectious Diseases*, 15(11):1273–1282, 2015.

[93] Edward PK Parker, Beate Kampmann, Gagandeep Kang, and Nicholas C Grassly. Influence of enteric infections on response to oral poliovirus vaccine: a systematic review and meta-analysis. *The Journal of infectious diseases*, 210(6):853–864, 2014.

[94] Liisa Piirainen, Mirja Stenvik, Merja Roivainen, Juhani Eskola, E Coen Beuvery, and Tapani Hovi. Randomised, controlled trial with the trypsin-modified inactivated poliovirus vaccine: assessment of intestinal immunity with live challenge virus. *Vaccine*, 17(9):1084–1090, 1999.

[95] Christopher D Pilcher, George Joaki, Irving F Hoffman, Francis EA Martinson, Clement Mapanje, Paul W Stewart, Kimberly A Powers, Shannon Galvin, David Chilongozi, Syze Gama, et al. Amplified transmission of HIV-1: comparison of HIV-1 concentrations in semen and blood during acute and chronic infection. *AIDS (London, England)*, 21(13):1723, 2007.

[96] Christopher D Pilcher, Hsiao Chuan Tien, Joseph J Eron, Pietro L Vernazza, Szu-Yun Leu, Paul W Stewart, Li-Ean Goh, and Myron S Cohen. Brief but efficient: acute HIV infection and the sexual transmission of HIV. *Journal of Infectious Diseases*, 189(10):1785–1792, 2004.

[97] Steven D Pinkerton. Probability of HIV transmission during acute infection in Rakai, Uganda. *AIDS and Behavior*, 12(5):677–684, 2008.

[98] Margarita Pons-Salort, Natalie A Molodecky, Kathleen M OReilly, Mufti Zubair Wadood, Rana M Safdar, Andrew Etsano, Rui Gama Vaz, Hamid Jafari, Nicholas C Grassly, and Isobel M Blake. Population immunity against serotype-2 poliomyelitis leading up to the global withdrawal of the oral poliovirus vaccine: Spatio-temporal modelling of surveillance data. *PLoS Med*, 13(10):e1002140, 2016.

[99] Kimberly A Porter, Ousmane M Diop, Cara C Burns, Rudolph H Tangermann, and SG Wassilak. Tracking progress toward polio eradication—worldwide, 2013–2014. *MMWR Morb Mortal Wkly Rep*, 64(15):415–20, 2015.

[100] Kimberly A Powers, Azra C Ghani, William C Miller, Irving F Hoffman, Audrey E Pettifor, Gift Kamanga, Francis EA Martinson, and Myron S Cohen. The role of acute and early HIV infection in the spread of HIV and implications for transmission prevention strategies in Lilongwe, Malawi: a modelling study. *The Lancet*, 378(9787):256–268, 2011.

[101] Kimberly A Powers, Mirjam E Kretzschmar, William C Miller, and Myron S Cohen. Impact of early-stage HIV transmission on treatment as prevention. *Proceedings of the National Academy of Sciences*, 111(45):15867–15868, 2014.

[102] Joseph Prejean, Ruiguang Song, Angela Hernandez, Rebecca Ziebell, Timothy Green, Frances Walker, Lillian S Lin, Qian An, Jonathan Mermin, Amy Lansky, et al. Estimated HIV incidence in the United States, 2006–2009. *PloS one*, 6(8):e17502, 2011.

[103] Nicoletta Previsani, Rudolph H Tangermann, Graham Tallis, and Hamid S Jafari. World Health Organization guidelines for containment of poliovirus following type-specific polio eradication—worldwide, 2015. *MMWR Morb Mortal Wkly Rep*, 64(33):913–7, 2015.

[104] Brandy L Rapatski, Frederick Suppe, and James A Yorke. HIV epidemics driven by late disease stage transmission. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 38(3):241–253, 2005.

[105] David A Rasmussen, Oliver Ratmann, and Katia Koelle. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS computational biology*, 7(8):e1002136, 2011.

[106] Leslie Roberts. New polio cases in Nigeria spur massive response. *Science*, 353(6301):738–738, 2016.

[107] EO Romero-Severson, E Volz, JS Koopman, T Leitner, and EL Ionides. Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men. *American journal of epidemiology*, 182(3):255–262, 2015.

[108] Ethan O Romero-Severson, Shah Jamal Alam, Erik M Volz, and James S Koopman. Heterogeneity in number and type of sexual contacts in a gay urban cohort. *Statistical communications in infectious diseases*, 4(1), 2012.

[109] Ethan O Romero-Severson, Grant D Meadors, and Erik M Volz. A generating function approach to HIV transmission with dynamic contact rates. *Mathematical modelling of natural phenomena*, 9(2):121–135, 2014.

[110] Ethan Obie Romero-Severson, Shah Jamal Alam, Erik Volz, and James Koopman. Acute-stage transmission of HIV: effect of volatile contact rates. *Epidemiology*, 24(4):516–521, 2013.

[111] Sarah E Rutstein, Jintanat Ananworanich, Sarah Fidler, Cheryl Johnson, Eduard J Sanders, Omar Sued, Asier Saez-Cirion, Christopher D Pilcher, Christophe Fraser, Myron S Cohen, et al. Clinical and public health implications of acute and early HIV detection and treatment: a scoping review. *Journal of the International AIDS Society*, 20(1), 2017.

[112] Lester M Shulman, Javier Martin, Danit Sofer, Cara C Burns, Yossi Manor, Musa Hindiyeh, Eugene Gavrilin, Thomas Wilton, Jacob Moran-Gilad, Ronni Gamzo, et al. Genetic analysis and characterization of wild poliovirus type 1 during sustained transmission in a population with >95% vaccine coverage, Israel 2013. *Clinical Infectious Diseases*, page ciu1136, 2014.

[113] Lester M Shulman, Ella Mendelson, Emilia Anis, Ravit Bassal, Michael Gdalevich, Musa Hindiyeh, Ehud Kaliner, Eran Kopel, Yossi Manor, Jacob Moran-Gilad, Daniella Ram, Danit Sofer, Eli Somekh, Diana Tasher, Merav Weil, Ronni Gamzu, and Itamar Grotto. Laboratory challenges in response to silent introduction and sustained transmission of wild poliovirus type 1 in Israel during 2013. *Journal of Infectious Diseases*, 210(suppl 1):S304–S314, 2014.

[114] LM Shulman, E Gavrilin, J Jorba, J Martin, CC Burns, Y Manor, J Moran-Gilad, D Sofer, MY Hindiyeh, R Gamzu, et al. Molecular epidemiology of silent introduction and sustained transmission of wild poliovirus type 1, Israel, 2013. *Euro Surveill*, 19(7):20709, 2014.

[115] Rudolf H Tangermann, Christine Lamoureux, Graham Tallis, and Ajay Goel. The critical role of acute flaccid paralysis surveillance in the Global Polio Eradication Initiative. *International Health*, 9(3):156–163, 2017.

[116] Kimberly M Thompson. Modeling poliovirus risks and the legacy of polio eradication. *Risk Analysis*, 33(4):505–515, 2013.

[117] Kimberly M Thompson and Radboud J Duintjer Tebbens. National choices related to inactivated poliovirus vaccine, innovation and the endgame of global polio eradication. *Expert review of vaccines*, 13(2):221–234, 2014.

[118] Kimberly M Thompson, Dominika A Kalkowska, and Radboud J Duintjer Tebbens. Managing population immunity to reduce or eliminate the risks of circulation following the importation of polioviruses. *Vaccine*, 33(13):1568–1577, 2015.

[119] Kimberly M Thompson, Mark A Pallansch, Radboud J Duintjer Tebbens, Steve G Wassilak, Jong-Hoon Kim, and Stephen L Cochi. Preeradication vaccine policy options for poliovirus infection and disease control. *Risk Analysis*, 33(4):516–543, 2013.

[120] Kimberly M Thompson, Mark A Pallansch, Radboud J Duintjer Tebbens, Steve G Wassilak, and Stephen L Cochi. Modeling population immunity to support efforts to end the transmission of live polioviruses. *Risk Analysis*, 33(4):647–663, 2013.

[121] Kimberly M Thompson, Peter M Strebel, Alya Dabbagh, Thomas Cherian, and Stephen L Cochi. Enabling implementation of the global vaccine action plan: Developing investment cases to achieve targets for measles and rubella prevention. *Vaccine*, 31:B149–B156, 2013.

[122] Kimberly M Thompson and Radboud J Duintjer Tebbens. Current polio global eradication and control policy options: perspectives from modeling and prerequisites for oral poliovirus vaccine cessation. *Expert review of vaccines*, 11(4):449–459, 2012.

[123] Kimberly M Thompson and Radboud J Duintjer Tebbens. Modeling the dynamics of oral poliovirus vaccine cessation. *Journal of Infectious Diseases*, 210(suppl 1):S475–S484, 2014.

[124] Lucia V Torian and Ellen W Wiewel. Continuity of HIV-related medical care, New York City, 2005–2009: Do patients who initiate care stay in care? *AIDS patient care and STDs*, 25(2):79–88, 2011.

[125] Theodore H Tulchinsky, Asad Ramlawi, Ziad Abdeen, Itamar Grotto, and Antoine Flahault. Polio lessons 2013: Israel, the West Bank, and Gaza. *The Lancet*, 382(9905):1611, 2013.

[126] Eric Vitinghoff, John Douglas, Frank Judon, David McKiman, Kate MacQueen, and Susan P Buchinder. Per-contact risk of human immunodeficiency virus transmission between male sexual partners. *American journal of epidemiology*, 150(3):306–311, 1999.

[127] Erik M Volz. Complex population dynamics and the coalescent under neutrality. *Genetics*, 190(1):187–201, 2012.

[128] Erik M Volz, Edward Ionides, Ethan O Romero-Severson, Mary-Grace Brandt, Eve Mokotoff, and James S Koopman. HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS medicine*, 10(12):e1001568, 2013.

[129] Erik M Volz, Sergei L Kosakovsky Pond, Melissa J Ward, Andrew J Leigh Brown, and Simon DW Frost. Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430, 2009.

[130] Bradley G Wagner, Matthew R Behrend, Daniel J Klein, Alexander M Upfill-Brown, Philip A Eckhoff, and Hao Hu. Quantifying the impact of expanded age group campaigns for polio eradication. *PloS one*, 9(12):e113538, 2014.

[131] Maria J Wawer, Ronald H Gray, Nelson K Sewankambo, David Serwadda, Xianbin Li, Oliver Laeyendecker, Noah Kiwanuka, Godfrey Kigozi, Mohammed Kiddugavu, Thomas Lutalo, et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *The Journal of Infectious Diseases*, 191(9):1403–1409, 2005.

[132] Xinyu Zhang, Lin Zhong, Ethan Romero-Severson, Shah Jamal Alam, Christopher J Henry, Erik M Volz, and James S Koopman. Episodic HIV risk behavior can greatly amplify HIV prevalence and the fraction of transmissions from acute HIV infection. *Statistical communications in infectious diseases*, 4(1), 2012.