Applications of Optimal Control Theory to Infectious Disease Modeling

by

Elsa Hansen

A thesis submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Doctor of Philosophy

Queen's University
Kingston, Ontario, Canada

January 2011

Abstract

This thesis investigates the optimal use of intervention strategies to mitigate the spread of infectious diseases. Three main problems are addressed:

(i) The optimal use vaccination and isolation resources under the assumption that these resources are limited. Specifically we address the problem of minimizing the outbreak size and we determine the optimal vaccination-only, isolation-only and mixed vaccination-isolation strategies.

(ii) The optimal use of a single antiviral drug to minimize the total outbreak size, under the assumption that treatment causes de novo resistance.

(iii) The optimal use of two antiviral drugs to minimize the total infectious burden. Specifically we address the situation where there are two different strains and each strain is effectively treated by only one drug.

Statement of Originality

The work in this thesis is predominantly my own. My supervisor Troy Day helped formulate the initial questions and also helped with some of the interpretations.

Acknowledgments

I would like to thank my supervisor Troy Day for teaching me how to ask questions and also for the constant support he has given me over the last few years. My environment was greatly enriched by Johanna, Amy and Helen who were always available to discuss math and have become dear friends. I would also like to thank Andrew Lewis for being my mentor for the first part of my graduate education.

Dedication

To my family for their endless support throughout the years.

To my dad Don, for all of the car rides, songs and conversations.

To my mom Jeanette, for being my champion and for staying up late with me when I was little to help me finish book reports.

To my brothers Max and Carl, for taking me on adventures from skipping rocks and crossing streams to climbing mountains.

To my little sister Jo, for letting me take her on adventures from twenty-five cent journeys across town to moonlit hikes down mountain sides.

To Tiffany and Tara for always making me feel like an important part of their families.

To Peter, Romy, Marley, Sasha and Axel, for reminding me what it is like to see the world for the first time.

To Kyle, for all the adventures we have had and will have. If I am with you then I am at home.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter we discuss the motivation for the work that follows, review our main analysis tool and detail the organization of the thesis.

1.1 Motivation

Mathematical modeling has become an important tool for assessing the effect of different intervention strategies on the spread of infectious diseases within a population. The majority of existing work falls into one of two categories. In the first category, intervention strategies are modeled by a constant parameter and the goal is to understand how changing the value of the parameter changes the dynamics of the system. Often the aim is to determine the best parameter value for a given performance measure. In the second category, intervention strategies are allowed to vary as a function of time and the goal is to determine the best function for a given performance measure. Due to the complexity of the models used, these studies are often numerical and are of varying generality. In some studies only a handful of specifically chosen functions are compared and in other studies a wide range of functions are considered. In the most general case, Pontryagin's Maximum Principle (PMP) can be used to compare a wide range of time varying functions. This is, for example, the approach taken in this thesis. The distinguishing feature between the majority of existing work that uses the PMP and the work found within these pages is that we focus on models that lend themselves to analytic solutions. Although the models are necessarily simpler, this approach has much to offer. For example our analysis provides results which do not depend on specific parameter values and therefore facilitates an understanding of the general mechanisms involved. Furthermore, our analysis often leads to analytic expressions that can be generalized to more detailed models.

We begin by reviewing our main analytic tool, Pontryagin's Maximum Principle.

1.2 Pontryagin's Maximum Principle

Pontryagin's Maximum Principle (PMP) is a classical result from optimal control theory that provides a necessary condition that must be satisfied by an optimal solution [35]. There are various versions of PMP for problem statements of varying generality. We present a version that is suitable for the types of problems discussed in this thesis; for more general versions see [13, 19].

Let $[t_0, t_f] \subset \mathbb{R}$, U be a bounded subset of $\mathbb{R}^m$ and $u : [t_0, t_f] \to U$ be a measurable function. For $i \in \{1, ..., n\}$ let $f = (f^1, ..., f^n) : \mathbb{R}^n \times U \to \mathbb{R}^n$ be continuous on U and continuously differentiable on $\mathbb{R}^n$. For a fixed u and constant $x_0 \in \mathbb{R}^n$ let $x : [t_0, t_f] \to \mathbb{R}^n$ be a function described by $x(t_0) = x_0$ and

$$\dot{x}^i(t) = f^i(x(t), u(t)) \quad \text{a.e on } [t_0, t_f],\ i \in \{1, ..., n\}. \tag{1.1}$$

We will call $u(t)$ the control applied at time t and $x : [t_0, t_f] \to \mathbb{R}^n$ the system trajectory corresponding to control u and initial condition x_0. Notice that in order to define a system trajectory a corresponding control and initial condition must be specified. We will call $x(t)$ the system state at time t.

Given a continuous function $\gamma : \mathbb{R}^n \to \mathbb{R}$ and a function $f^0 : \mathbb{R}^n \times U \to \mathbb{R}$ that is continuous on U and continuously differentiable on $\mathbb{R}^n$ define the cost function

$$J(x, u) = \int_{t_0}^{t_f} f^0(x(t), u(t))\mathrm{d}t + \gamma(x(t_f)).$$

Roughly speaking, the goal is to find a control $u : [t_0, t_f] \to U$ and corresponding trajectory $x : [t_0, t_f] \to \mathbb{R}^n$ that minimizes J. There are different variations of this problem depending on the specific constraints placed on t_0, t_f, $x(t_0)$ and $x(t_f)$. Before giving a formal problem statement, we first review some terminology.

Definition. Let $I_0,\ I_f \subset \mathbb{R}$ denote the sets of possible values for the initial time t_0 and the final time t_f, respectively. Let $S_0,\ S_f \subset \mathbb{R}^n$ denote the sets of possible values for the initial and final state values, respectively. Then **the set of allowable boundary values for a trajectory** is defined by $\mathcal{B} = \{(t_0, x_0, t_f, x_f) \mid t_0 \in I_0,\ x_0 \in S_0,\ t_f \in I_f \text{ and } x_f \in S_f\}$.

Assumption 1.1. We will assume that there exist functions $\Psi_0,\ \Psi_f \in C^1(\mathbb{R}^n, \mathbb{R})$ such that $S_0 = \Psi_0^{-1}(0)$ and $S_f = \Psi_f^{-1}(0)$ and $D\Psi_0(z)$ and $D\Psi_f(y)$ are surjective for all $z \in S_0$ and $y \in S_f$. Here, $D\Psi_i$ denotes the Jacobian of Ψ_i.

Definition. Denote by $U \subset \mathbb{R}^m$ the bounded set of all allowable values for the control. Then an **admissible control** is a measurable function $u : [t_0, t_f] \to U$ that satisfies:

(i) there exists an absolutely continuous function $x : [t_0, t_f] \to \mathbb{R}^n$,

(ii) $\dot{x}(t) = f(x(t), u(t))$ a.e on $[t_0, t_f]$,

(iii) $t \mapsto f^0(x(t), u(t))$ is integrable on $[t_0, t_f]$ and

(iv) $(t_0, x(t_0), t_f, x(t_f)) \in \mathcal{B}$.

The trajectory x corresponding to an admissible control is an **admissible trajectory**. The corresponding pair (x, u) is an **admissible pair**.

Definition. For any $\hat{b} = (b_0, b_1, ..., b_n) \in \mathbb{R}^{n+1}$ there is a corresponding value $b = (b_1, ...b_n) \in \mathbb{R}^n$. Define $\hat{f} : \mathbb{R}^{n+1} \times U \to \mathbb{R}^{n+1}$ by $\hat{f}^i(\hat{b}, u) = f^i(b, u)$ for all $i \in \{0, 1, ..., n\}$. Now define the **extended system of differential equations** by

$$\begin{aligned} \dot{\hat{x}}^0 &= \hat{f}^0(\hat{x}, u), \\ \dot{\hat{x}}^j &= \hat{f}^j(\hat{x}, u), \quad j \in \{1, ..., n\}. \end{aligned}$$

Definition. The **extended Hamiltonian** is the function $H : \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \times U \to \mathbb{R}$ defined by

$$H(\hat{a}, \hat{b}, c) = \sum_{k=0}^{n} \hat{a}_k \hat{f}^k(\hat{b}, c).$$

Remark 1.1. In the sequel D_2H will denote the derivative of H with respect to its second argument.

Problem Statement 1.1. Assume U is a bounded set in $\mathbb{R}^m$. Assume $\hat{f}$ is continuous on U and continuously differentiable on $\mathbb{R}^{n+1}$. Let $\mathcal{A}$ denote the set of all admissible pairs (x, u) and let $\mathcal{A}$ be non-empty. Find an admissible pair $(x^*, u^*) \in \mathcal{A}$ such that $J(x^*, u^*) \leq J(x, u)$ for every $(x, u) \in \mathcal{A}$.

Theorem 1.1. *(Pontryagin's Maximum Principle)*

If (x^*, u^*) is a solution to Problem Statement 1.1 then there exists an absolutely continuous function $\hat{\lambda} : [t_0, t_f] \to \mathbb{R}^{n+1}$ such that

(i) $\dot{\hat{\lambda}} = -D_2 H(\hat{\lambda}, \hat{x}, u)$, *a.e on* $[t_0, t_f]$,

(ii) $\hat{\lambda}(t) \neq 0$ *for all* $t \in [t_0, t_f]$,

(iii) $\hat{\lambda}_0 \in \{0, -1\}$,

(iv) $H(\hat{\lambda}(t), \hat{x}^*(t), u^*(t)) = \sup_{u \in U} H(\hat{\lambda}(t), \hat{x}^*(t), u)$, *a.e on* $[t_0, t_f]$,

(v) there exists $c \in \mathbb{R}$ *such that* $H(\hat{\lambda}(t), \hat{x}^*(t), u^*(t)) = c$, *a.e on* $[t_0, t_f]$,

(vi) if the end time t_f *is free then* c *can be taken to be zero, and*

(vii) $\lambda(t_0)$ *is orthogonal to* $\ker(D\Psi_0(x^*(t_0)))$ *and* $\lambda(t_f)$ *is orthogonal to* $\ker(D\Psi_1(x^*(t_f)))$.

Remark 1.2. Part (vii) of PMP is often referred to as the transversality condition and it states that $\lambda(t_0)$ and $\lambda(t_f)$ must be perpendicular to the constraint sets S_0 and S_f respectively.

An important subtlety of PMP is that it presupposes that an optimal control exists. Therefore, before applying the PMP the first step is to prove that the problem under consideration actually possesses a solution. The following theorem from [19] can be used to prove the existence of a solution for a number of optimal control problems.

Assumption 1.2. Let U be a bounded subset of $\mathbb{R}^m$ and assume $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ is continuous on both $\mathbb{R}^n$ and U. Furthermore, assume that there exist constants C_1 and C_2 such that

(i) $\|f(x,u)\| \leq C_1(1+\|x\|)$ and

(ii) $\|f(\tilde{x},u) - f(x,u)\| \leq C_2\|\tilde{x} - x\|$ for all $\tilde{x}$, $x \in \mathbb{R}^n$ and $u \in U$.

Theorem 1.2. *Suppose Assumption 1.2 holds. Futhermore, assume*

(i) $f^0 : \mathbb{R}^n \times U \to \mathbb{R}$ *is continuous,*

(ii) the set of admissible pairs $\mathcal{A}$ *is not empty,*

(iii) U *is compact,*

(iv) γ *is continuous on* $\mathcal{B}$*,*

(v) the set of allowable boundary values $\mathcal{B}$ *is compact and*

(vi) $\tilde{F}(x) = \{(z_0, z) \parallel z_0 \geq f^0(x,u), z = f(x,u), u \in U\}$ *is convex.*

Then there exists a solution to Problem 1.1.

1.3 Organization of Thesis

This thesis is comprised of four self-contained chapters. Chapter 2 uses a number of simplifying asumptions in order to present a simple geometric picture of the complicated problem of treatment administration. Included in this discussion is how (i) limited resources and (ii) de novo resistance change the treatment strategy that minimizes the total infectious burden. This chapter has been submitted for publication in the Canadian Applied Math Quarterly Focus Issue on Pandemic Modeling whose targeted audience includes public health surveillance experts and decision makers. Chapter 2 also serves as an introdution to Chapter 3 and Chapter 4. Chapter 3 makes a detailed study of optimal vaccination and isolation strategies under the assumption that there are limited resources. This chapter has been published previously in the Journal of Mathematical Biology [23]. Chapter 4 addresses the optimal use of a single antiviral drug given that treatment can cause de novo resistance. This chapter has been previously published in Proceedings of the Royal Society B [24]. Finally, Chapter 5 generalizes the problem of optimal treatment administration with de novo resistance to the case when there are two different drugs.

Chapter 2

Optimal Treatment: A Geometric Picture

2.1 Abstract

Despite much debate and numerous studies, the question of how to most effectively administer antiviral treatment during an influenza pandemic is still unanswered. We present some basic models, coupled with simple geometric arguments, to provide insight into this topic. Specifically, we address the question of how best to administer treatment to minimize the outbreak size. Furthermore, we consider how a limited supply of drugs and the occurrence of de novo resistance can affect this solution.

2.2 Introduction

An important and unresolved issue in health policy is how to most effectively administer antiviral treatment during an influenza pandemic. This is a very complicated topic that involves many confounding factors and has been the focus of many modeling efforts [for a sample of existing work see 5, 6, 9, 10, 14, 18, 22–24, 28, 30–33, 36, 38–40, 43]. There is in fact so much information on the topic that it is easy to become overwhelmed. Here, we take a step back and discuss how some basic models, coupled with simple geometric arguments, can provide insight into the effects of treatment. Specifically, we address the question of how best to administer treatment and consider how limited resources and the occurrence of de novo resistance can affect this solution. We do not intend to offer a complete solution to the problem of treatment administration but rather our aim is to isolate some important relationships and provide simple intuition with which more complicated scenarios can be approached.

2.3 Basic Model

The first step in solving for the best treatment strategy is to define what is meant by the "best" or "optimal" treatment strategy. For the purposes of this analysis, the "optimal" treatment strategy is the one that minimizes the attack ratio

$$A = \frac{S_0 - S_f}{S_0},$$

where S_0 and S_f are the number of susceptibles at the start and end of the outbreak, respectively. Although minimizing the attack ratio is a reasonable goal, it should be emphasized that the attack ratio is not the only outcome of an epidemic that policy makers strive to control.

The starting point of our analysis will be the following simple model of an outbreak:

$$\begin{aligned} \dot{S} &= -\beta IS, \\ \dot{I} &= \beta IS - \mu I. \end{aligned} \tag{2.1}$$

In Model (2.1) the rate at which susceptibles (S) become infected is determined by the mass action law (with transmission constant β) and infecteds (I) either die or recover with immunity at a constant per capita rate (μ). The motivation for using such a simple model is that it allows us to isolate the mechanisms involved in treatment and to measure how systematically changing the treatment assumptions changes the outcomes. It is also worthwhile to note that if the initial number of infecteds (I_0) and susceptibles (S_0) are fixed then, for Model (2.1), minimizing the attack ratio is equivalent to (i) minimizing the total infectious burden, (ii) minimizing the total outbreak size and (iii) maximizing the final number of susceptibles. In the sequel these outcomes will be used interchangeably.

Model (2.1) separates the population into three groups: susceptible individuals, infected individuals and removed individuals. Removed individuals are those who have either died or recovered with immunity and hence are no longer involved in the spread of the disease. The course of the outbreak therefore, is determined by how the number of susceptibles and infecteds change.

Figure 2.1 shows that the number of susceptibles is a decreasing function of time (Panel A) and that the number of infected individuals initially increases, reaches

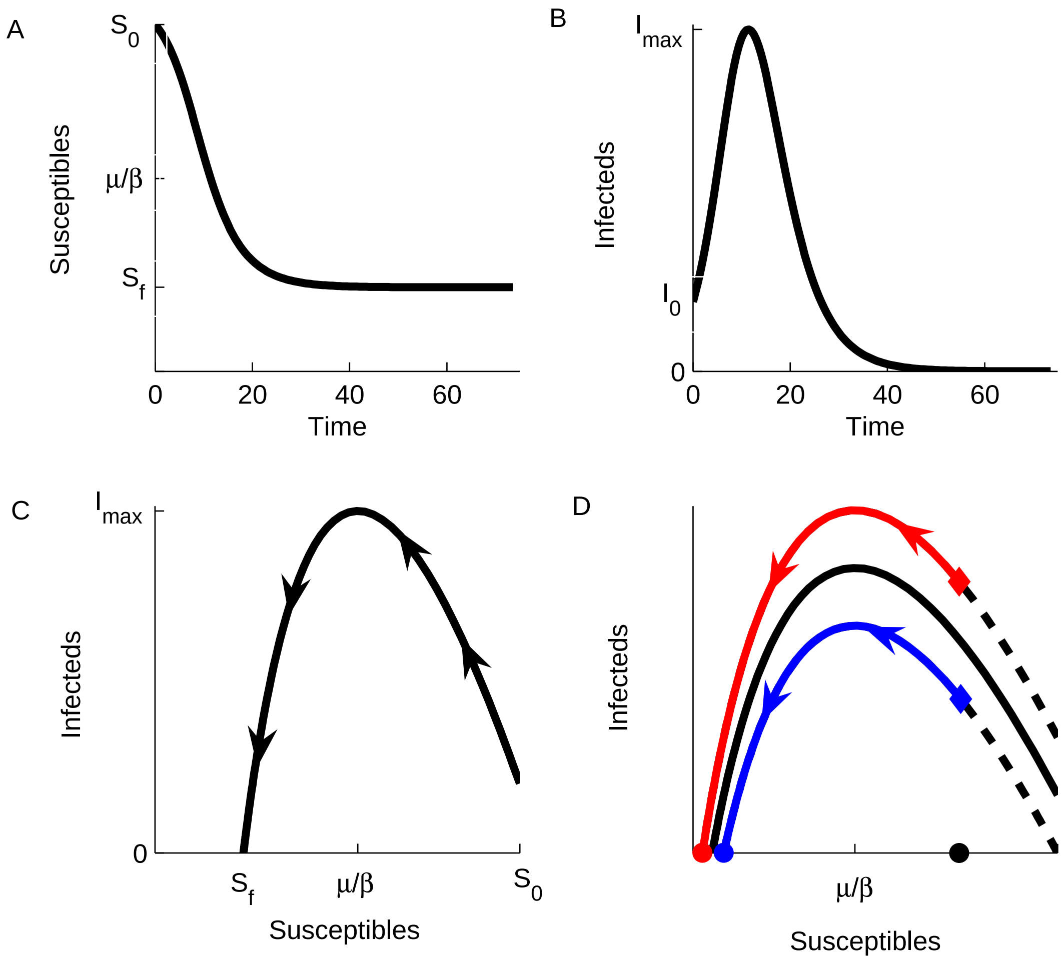

Figure 2.1: Outbreak dynamics for Model (2.1). **Panel A**: Shows the number of susceptible individuals as a function of time. **Panel B**: Shows the number of infected individuals as a function of time. **Panel C**: Shows the number of infected individuals as a function of the number of susceptible individuals. The outbreak flows in the direction of the arrows. **Panel D**: Each curve is similar to the curve in Panel C except for a different set of initial conditions (S_0, I_0). The red curve shows the path of a possible outbreak whose start (end) is marked by the red diamond (circle). The blue curve shows the path of another outbreak with different initial conditions. Because the blue curve lies below the red curve it has a lower attack rate. If at the start of the red outbreak all of the infected individuals are successfully treated then the number of infecteds will jump to zero (black circle) and the outbreak will immediately end.

a maximum $I_{\max}$ and then decreases until the outbreak ends (Panel B). Together, Panel A and Panel B completely specify the course of the outbreak through time. If however, we are not interested in the specific time that events occur then Panel A and Panel B can be combined into a single figure (Panel C). Panel C shows how the number of infecteds changes as a function of the number of susceptibles. Because the number of susceptibles is a decreasing function of time (Panel A) the course of the outbreak moves from right to left in Panel C (i.e., the outbreak flows in the direction of the arrows). The plot in Panel C is called a phase portrait and can be particularly useful because it highlights dynamical relationships between the states of the system. For example, Panel C shows that the number of infecteds reaches its maximum $I_{\max}$ precisely when $S = \frac{\mu}{\beta}$. All of the graphical interpretations that we will discuss will involve phase portraits. In fact we can use the phase portrait in Panel D to gain some key insights into what an optimal treatment strategy should do. Each curve in Panel D plots the number of infected individuals as a function of the number of susceptibles for a given set of initial conditions (i.e., each curve in Panel D represents the same information as the curve in Panel C, but for different values of S_0 and I_0). For example, if the red diamond corresponds to the start of an outbreak then from this starting point, the outbreak will follow the red curve until the end of the outbreak, which is marked by the red circle. If on the other hand, the start of the outbreak corresponds to the blue diamond then the outbreak will follow the blue curve until the end of the outbreak (blue circle). An important observation is that each curve in Panel D can be obtained by vertically translating the curve below it. Consequently, for any two outbreak paths, the lower path will have a lower outbreak size. For example, in Panel D the blue path lies below the red path and correspondingly the blue circle lies to the right of the red circle (i.e., when the outbreak corresponding to the blue path ends there are more susceptibles remaining than when the outbreak corresponding to the red path ends). Therefore, at least intuitively speaking, the optimal treatment strategy is the strategy that forces the outbreak to follow the lowest possible curve.

2.4 Modeling Treatment

Perhaps the simplest way to model treatment is to assume that (i) treatment can be administered instantly and (ii) treatment completely prevents transmission. In the context of Model (2.1) this means that at any time during the outbreak any number of infected individuals can be treated and these treated individuals instantly stop transmitting the disease (hence they essentially enter the "removed" class). Under these assumptions, clearly the best strategy is to immediately treat all infected individuals. By doing this, all of the infected individuals will immediately stop transmitting the disease and the outbreak will end. For instance, if the start of an outbreak is marked by the red diamond in Figure 2.1 (Panel D), then immediately treating all of the infected individuals means that instead of following the red path, the outbreak would immediately jump to the black circle (where $I = 0$) and the outbreak would end with $S_0 = S_f$.

The treatment model just described is highly idealized but nonetheless highlights some fundamental mechanisms that will feature prominently in the rest of our analysis. In the following two sections we employ this idealized model to present some simple geometric arguments that describe the optimal treatment strategy when (i) there is a limited supply of antiviral drugs and (ii) when there is a possibility that treatment can lead to de novo resistance. We also discuss how removing the assumptions that (i) treatment can be administered instantly and that (ii) treatment completely prevents transmission changes the conclusions of these geometric arguments.

2.4.1 Optimal Treatment when Resources are Limited

Two important aspects of pandemic planning are (i) deciding the amount of antiviral drugs to stockpile in preparation for a future outbreak and (ii) deciding how to administer a stockpile of antivirals once an outbreak begins. In an actual outbreak, optimal distribution of antivirals is especially relevant when there are not enough doses to treat everyone that becomes infected during the outbreak. For our idealized treatment model a stockpile of drugs is insufficient if there are not enough drugs to treat all infected individuals at the start of the outbreak. For example if there is a delay in detecting the start of the outbreak, the number of infections can become quite large before treatment begins and hence there may not be enough resources to treat all infected individuals. In this case, the optimal treatment strategy (Result 2.1) can be determined using a simple geometrical argument (Box A).

Result 2.1. *Suppose (i) treatment can be administered instantly and (ii) treated individuals cannot transmit infection. Furthermore, suppose that during the course of the entire epidemic at most $T_{\max}$ individuals can be treated. If $I_0 \leq T_{\max}$ then the optimal treatment strategy for Model* (2.1) *is to immediately treat all infected individuals. If, on the other hand, there are insufficient resources to treat all of the infected individuals at the start of the outbreak ($I_0 > T_{\max}$) then any treatment strategy that uses all of the available resources is optimal.*

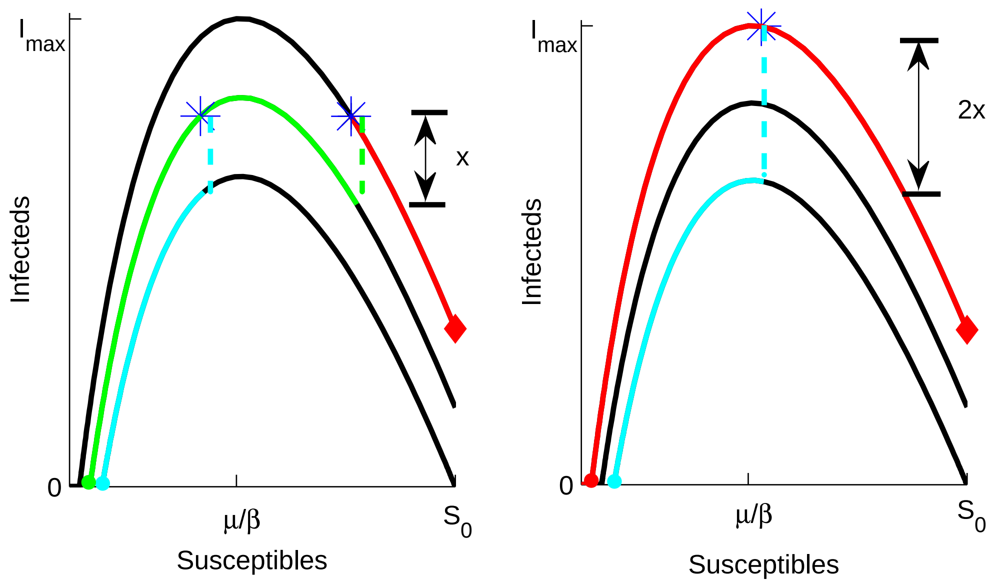

Figure 2.2: Effect of different treatment strategies on dynamics of Model (2.1): **Panel A**: The red diamond marks the start of an outbreak. If there is a single treatment effort that successfully treats x individuals (right most blue star) then the end of the outbreak is marked by the green circle. If there is a second treatment effort that successfully treats x individuals (left most blue star) then the end of the outbreak is marked by the cyan circle. The "double" treatment effort results in a lower attack rate than the "single" treatment strategy (the cyan circle lies to the right of the green circle). **Panel B**: The red diamond marks the start of an outbreak. If there is a single treatment effort that successfully treats $2x$ individuals (blue star) then the end of the outbreak is marked by the cyan circle. Thus the "double" treatment effort that treats $2x$ individuals (Panel A) results in the same attack rate at the "single" treatment effort that treats $2x$ individuals (Panel B).

Box A: Geometric Interpretation of Result 2.1

The possible outbreak paths of Model (2.1) are vertical translations of each other and treatment allows an outbreak to "jump" to lower paths. Therefore, since the final number of susceptibles is determined by which curve the outbreak ultimately follows, once an outbreak begins all that matters is how many individuals are treated and not how treatment is distributed throughout time.

Figure 2.2 provides a graphical explanation of this result. Each panel in Figure 2.2 demonstrates the effects of a different treatment strategy on an outbreak. Although the treatment strategies differ, they both treat the same number of individuals and hence result in the same attack ratio. The treatment strategy in Panel A involves two isolated treatment efforts, with each effort treating x individuals (the location of these treatment efforts are marked by blue stars). Once the outbreak starts (red diamond), the outbreak follows the red curve until the first treatment effort occurs (right most blue star). Since treated individuals do not transmit the infection these x individuals are removed and the number of infections instantly jumps from I_1 to $I_1 - x$. Now the outbreak will continue along the green curve. If no more treatment occurs then the outbreak will continue along the green curve until the end of the epidemic (green circle). If, on the other hand, an additional x individuals are treated (left most blue star) then the number of infecteds will instantly jump from I_2 to $I_2 - x$. Now the outbreak will continue along the cyan curve until the end of the outbreak (cyan circle). Notice that a single treatment effort (treating x individuals) results in an outbreak with a higher attack ratio then two treatment efforts (treating a total of $2x$ individuals). The outbreak ending at the cyan circle has a lower attack rate than the one ending at the green circle not because two treatment efforts occurred but rather because in total more individuals were treated. To emphasize this fact, Panel B shows that if $2x$ individuals are treated during a single treatment effort then the outbreak will also ultimately follow the cyan curve and hence the final number of susceptibles is the same as for the "double effort" treatment strategy in Panel A.

Result 2.1 is particularly interesting because it holds even in less idealized situations. Consider, for example, the following modification of Model (2.1):

$$\begin{aligned}\dot{S} &= -\beta(I+\delta_T T)S,\\ \dot{I} &= \beta(I+\delta_T T)S-(\mu+u)I,\\ \dot{T} &= uI-\mu T,\end{aligned} \tag{2.2}$$

where treated individuals are denoted by T, $\delta_T \in [0,1)$ represents the decrease in transmissibility due to treatment, $u \in [0, u_{\max}]$ is the per capita treatment rate and u is allowed to vary in time. In Model (2.2), an infected individual will either eventually die, eventually recover with immunity or eventually receive treatment and continue to transmit the infection at a reduced rate.

Model (2.2) does not use the two assumptions that made the geometrical interpretation of Result 2.1 possible. In Model (2.2) treatment is not administered instantly but rather at the bounded rate uI. Among other things, this ensures that an individual will spend some time circulating in the population spreading the disease before they are treated. Furthermore, provided $\delta_T > 0$, a treated individual will still transmit the disease. In Chapter 3 it is shown that if $\delta_T = 0$ and there are insufficient

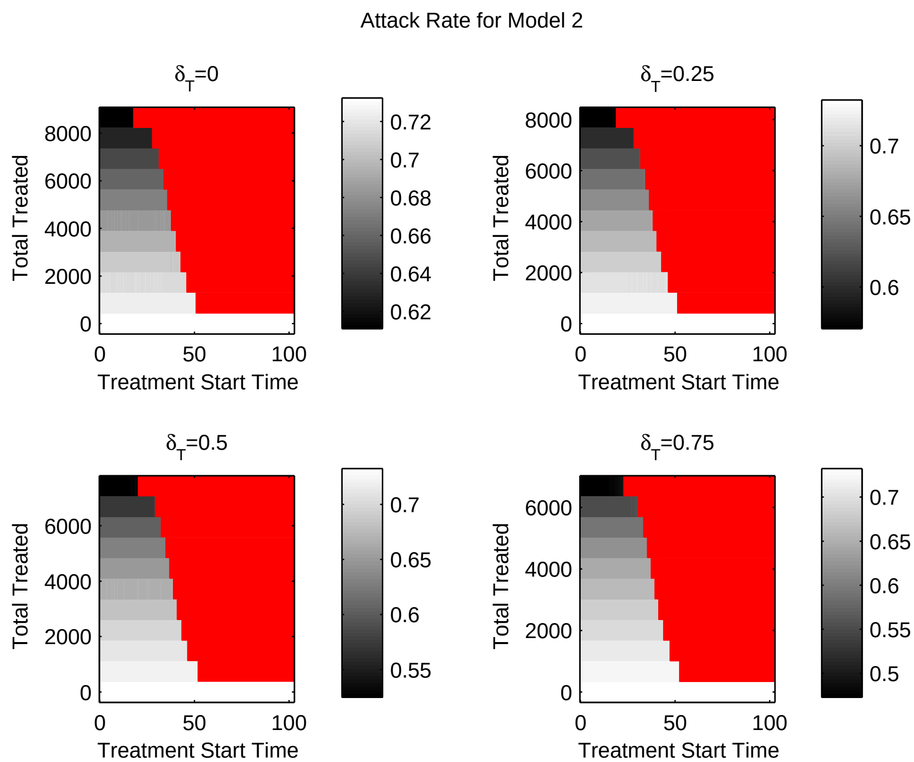

Figure 2.3: Attack rate as a function of the treatment start time and the total number of treated individuals for Model (2.2). Each panel corresponds to a different δ_T (if $\delta_T = 0$ then treated individuals do not transmit the disease). Each point on a panel indicates the attack rate that occurs when treatment starts at the time indicated on the horizontal axis and continues at the maximal rate until the total number of treated individuals reaches the value indicated on the vertical axis. Once the number of treated individuals is equal to the number indicated on the vertical axis treatment stops (u is set to 0) and the outbreak continues without any intervention until it ends. The red areas indicate start times for which it is not possible to treat the number of individuals indicated on the vertical axis. Hence the red areas should be ignored. Furthermore, if we let $T_{\max}$ denote the number of individuals treated when treatment starts immediately and continues at the maximum rate for the entire outbreak. Then we only consider treatment strategies that result in fewer than T_{max} treated individuals (the reason for this will be detailed in the Discussion). Increasing the total number of treated individuals decreases the attack rate (the grey scale darkens as we move from the bottom to the top of a panel). The start time does not affect the attack rate (the grey scale is relatively constant as we move from left to right). Hence it appears that the attack rate is determined by the total number of treated individuals and not how these individuals are treated throughout time. Parameters are $\mu = \frac{1}{3.3}, \frac{\beta}{\mu} = 1.8$, $S_0 = 100000$, $I_0 = 10$, $T(t_0) = 0$, $u_{\max} = 0.05$.

resources to treat with maximal effort for the entire outbreak then any strategy that uses all of the available resources is optimal. Figure 2.3 suggests that this result is also true when $\delta_T > 0$.

2.4.2 Optimal Treatment when there is Treatment Resistance

An important issue that has recently garnered a lot of attention is the de novo appearance of treatment-resistant strains [5, 14, 18, 22, 24, 28, 30, 32, 33, 36, 38–40, 43]. If treatment can cause resistance then treatment may have a negative effect on the outcome of an outbreak. In fact, for our idealized treatment model, it may no longer be optimal to immediately treat all infected individuals. This is because, individuals acquiring de novo resistance can initiate a second treatment-resistant outbreak. Thus, even though treating immediately will control the original treatment-sensitive outbreak, an outbreak may still occur.

Suppose the probability that a treated individual develops resistance is κ. Furthermore, suppose that if an individual develops resistance de novo then it does so immediately after being treated. Then if at $t = \tau$ all infected individuals are treated, $\kappa I(\tau)$ individuals will become resistant and $(1-\kappa)I(\tau)$ individuals will be successfully treated. Using our idealized treatment model, treatment will stop the original treatment-sensitive outbreak and will "seed" a new treatment-resistant outbreak with $\kappa I(\tau)$ resistant infections. Therefore an outbreak with de novo resistance can be modeled by the following sets of differential equations:

Table 2.1: A model for idealized treatment that can cause de novo resistance

	Before Treatment	After Treatment
Model	$\dot{S} = -\beta SI$ $\dot{I} = \beta SI - \mu I$	$\dot{S} = -\delta_R \beta SR$ $\dot{R} = \delta_R \beta SR - \mu R$
Initial Conditions	$(S(t_0), I(t_0)) = (S_0, I_0)$	$(S(\tau), R(\tau)) = (S(\tau), \kappa I(\tau))$

In the second set of equations R denotes resistant infections and we assume that resistance confers a fitness cost that reduces the transmission constant from β to $\delta_R \beta$, where $0 < \delta_R < 1$. If $\kappa I_0 < 1$ then treating immediately will create fewer than one resistant individual with which to "seed" the resistant outbreak. In this situation treating immediately essentially ends the outbreak. If on the other hand $\kappa I_0 > 1$ then treating immediately can cause a treatment-resistant outbreak. In this case immediate treatment is no longer the obvious best choice, and a reasonable question to ask is whether or not it is beneficial to delay treatment. For our idealized treatment model, the optimal amount of delay is described in Result 2.2.

Result 2.2. *Suppose (i) treatment can be administered instantly and (ii) treated individuals cannot transmit infection. Furthermore, suppose that (i) treated individuals develop resistance with probability* κ $(0 < \kappa < 1)$*, (ii) if a treated individual does develop resistance it does so immediately after treatment and (iii) the transmission coefficient of the resistant strain is* $\delta_R \beta$ $(0 < \delta_R \leq 1)$*. If* $\kappa I_0 \leq 1$ *then it is optimal to treat immediately and thus immediately end the outbreak. If* $\kappa I_0 > 1$ *then it is optimal to delay treatment until* $S = S^* = \frac{\mu}{\delta_R \beta}(1 + \kappa \frac{1-\delta_R}{1-\kappa})$*. If* $S_0 < S^*$ *then treatment should* *start immediately.*

Unlike Result 2.1 which describes a number of optimal treatment strategies, Result 2.2 specifies a single optimal treatment strategy. There are two reasons for this difference. First, in Result 2.1 the total number of treated individuals is determined by the stockpile size whereas in Result 2.2 it is determined by the number of infected individuals at the time of treatment. More specifically, in Result 2.1 the maximum total number of treated individuals is $T_{\max}$ whereas in Result 2.2 the total number of treated individuals is a fixed fraction of $I(\tau)$ (i.e., Total Treated $= \kappa I(\tau)$, where τ is the treatment start time). Secondly, in Result 2.1 the path that the outbreak follows after treatment is a vertical translation of the path that it followed before treatment. On the other hand, for Result 2.2 the path that the outbreak follows after treatment is determined by the characteristics of the resistant strain and this path may be completely different from the path it followed before treatment (since the original path is determined by the treatment-sensitive strain). Nevertheless, Result 2.2 can also be understood using a simple geometrical argument (Box B).

Box B: Geometric Interpretation of Result 2.2

Suppose that immediately treating all infected individuals does not immediately end the outbreak. In this case the optimal treatment strategy is the strategy that forces the outbreak to eventually follow the lowest possible resistant outbreak path. Result 2.2 says that the outbreak will eventually follow the lowest possible resistant outbreak path if the treatment effort occurs when the sensitive outbreak path (scaled by the probability of de novo resistance) is tangent to the resistant outbreak path.

Figure 2.4 provides a graphical explanation of this result. Panel A illustrates the case when $\delta_R = 1$. In this case the outbreak paths of the resistant strain (blue curves) are vertical translations of the outbreak path of the sensitive strain (black curve). Therefore the optimal treatment strategy is the one that treats the most individuals (by treating the maximum number of individuals the outbreak is able to jump to the lowest possible curve). Since the number of treated individuals is $\kappa I(\tau)$ it is optimal to treat when the number of sensitive infections has reached a maximum (since $\kappa I(\tau) \leq \kappa I_{\max}$). Therefore, it is optimal to treat when $S = \frac{\mu}{\beta}$ (indicated by the left most arrow, Panel A). If treatment occurs at any other time fewer individuals will be treated and so the outbreak will ultimately follow a higher path (such a scenario is indicated by the right most arrow, Panel A).

Panels B, C and D illustrate the case when $0 < \delta_R < 1$. In this case the resistant outbreak paths (blue curves, Panel B) are not vertical translations of the sensitive outbreak path (solid black curve, Panel B). Now, since the resistant outbreak is initiated by $\kappa I(\tau)$ resistant individuals, multiplying the original sensitive outbreak curve by κ shows the possible initial conditions of the resistant outbreak. The dashed black curve in Panel C is the original sensitive outbreak curve (solid black curve, Panel B) multiplied by κ. The blue curves in Panel C show possible resistant outbreaks that originate from the κI curve (dashed black curve). Panel D is a close up of Panel C. The point where the dashed red line intersects the black dashed curve indicates where the κI curve is tangent to the resistant outbreak curve that originates from the κI curve. We will call the resistant outbreak curve that originates from the intersection of the κI curve and the dashed red line the "optimal resistant outbreak curve". Any resistant outbreak curve that originates from the right of the dashed red line has an initial slope that is larger in magnitude than the slope of the κI curve. Conversely, any resistant outbreak curve that originates from the left of the red dashed line has an initial slope that is smaller in magnitude than the slope of the κI curve. Clearly, any resistant outbreak that initiates from the right of the vertical line will have an endpoint that lies to the left of the endpoint of the optimal resistant outbreak curve (and hence a larger attack rate). Similarly, (since the outbreak curves never cross) any resistant outbreak that initiates from the left of the red dashed line will have an endpoint that lies to the left of the endpoint of the optimal resistant outbreak curve (and hence a larger attack rate). Therefore, the optimal time to start treatment is when the κI curve and the resistant outbreak originating from the κI curve are tangent.

From the equations for the sensitive and resistant models we have:

$$\kappa \frac{\mathrm{d}I}{\mathrm{d}S} = \frac{\kappa I(\beta S - \mu)}{-\beta I S}, \tag{2.3}$$

and

$$\frac{\mathrm{d}R}{\mathrm{d}S} = \frac{R(\delta_R \beta S - \mu)}{-\delta_R \beta R S}. \tag{2.4}$$

Setting Equation (2.3) and Equation (2.4) equal to each other and solving for S gives:

$$S = \frac{\mu}{\delta_R \beta}(1 + \kappa \frac{1 - \delta_R}{1 - \kappa}) = S^*. \tag{2.5}$$

Notice that when $\delta_R = 1$ Equation (2.5) simplifies to $S = \frac{\mu}{\beta} = S^*$ and this coincides with the analysis of Panel A.

Result 2.2 is particularly interesting because similar results are true in less idealized situations. Consider, for example, the following modification of Model (2.1):

$$\begin{aligned}
\dot{S} &= -\beta(I + \delta_T T + \delta_R R)S, \\
\dot{I} &= \beta(I + \delta_T T)S - (\mu + u)I, \\
\dot{T} &= uI - (\mu + \nu)T, \\
\dot{R} &= \beta\delta_R RS + \nu T - \mu R,
\end{aligned} \tag{2.6}$$

where ν is the per capita rate at which treated individuals develop resistance and all other parameters are as previously defined. In this Model, an infected individual will either eventually die, eventually recover with immunity or eventually receive treatment. A treated individual will continue to transmit the infection (at a reduced rate) and will eventually die, eventually recover with immunity or eventually develop resistance. Finally, a resistant individual will transmit a treatment-resistant strain and will eventually die or recover with immunity.

Similar to Model (2.2), Model (2.6) does not use the two assumptions that made the geometrical interpretation possible. Although a complete analysis of this model is beyond the scope of this dissertation, if $\delta_T = 0$ and de novo resistance occurs immediately after treatment then it is proved in Chapter 4 that (i) it may be optimal to delay the start of treatment and (ii) once treatment begins it is best to treat at the maximal rate (i.e., $u = u_{\max}$). Furthermore, it is shown in Chapter 4 that the quantity S^* (defined in Result 2.2) provides an upper bound for the optimal amount of delay. Namely, treatment should never be delayed longer than the time it takes for

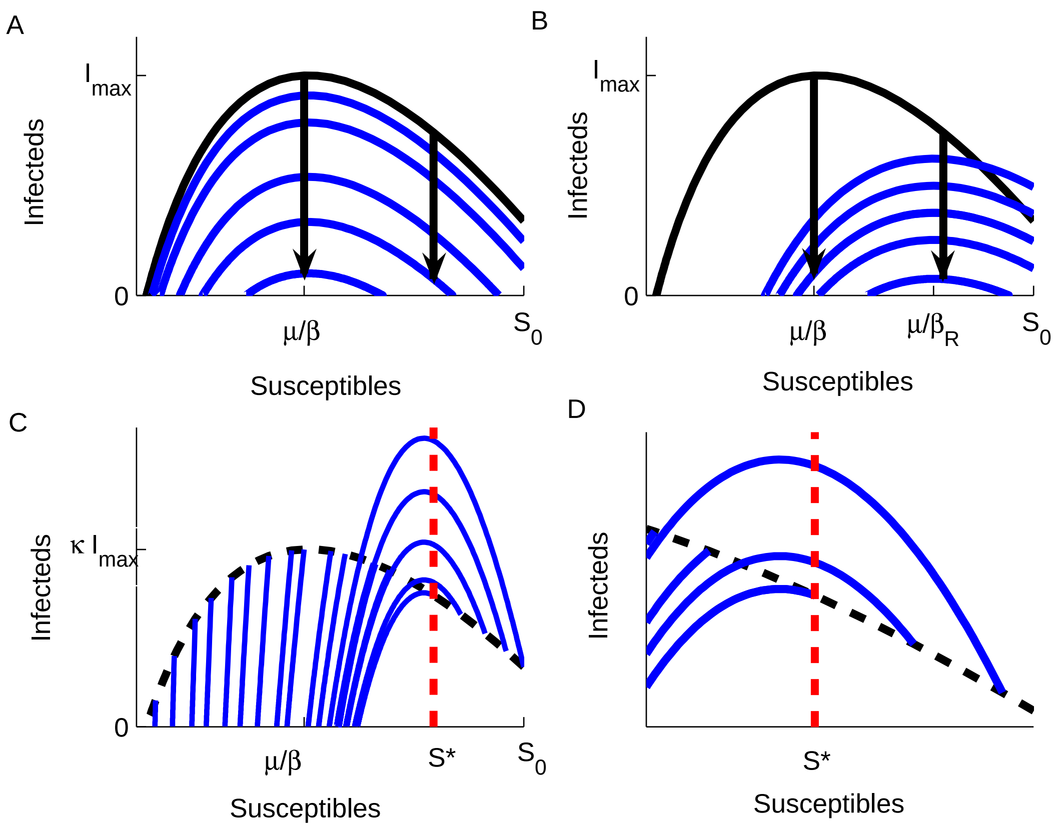

Figure 2.4: Effect of different treatment start times on the dynamics of the model described in Table 2.1: **Panel A**: If there is no fitness cost for resistance then the possible resistant outbreak paths (blue curves) are vertical translations of the sensitive outbreak path (black curve). The optimal treatment strategy treats the most individuals (left most black arrow) and thus forces the outbreak to follow the lowest possible resistant outbreak path. Treating fewer individuals (right most arrow) will result in a higher attack rate. **Panel B**: If there is a fitness cost for resistance then the possible resistant outbreak paths (blue curves) are **not** vertical translations of the sensitive outbreak path (black curve). Treating the maximum number of individuals (left most arrow) is no longer the obvious best choice. In fact, treating fewer individuals (right most arrow) can result in a lower attack rate. **Panel C**: The dashed black curve is κ times the black curve in Panel B and indicates the possible initial conditions for the resistant outbreak. Different initial conditions result in different resistant outbreak paths (blue curves). There is a single optimal resistant outbreak curve. The initial conditions of this "optimal resistant outbreak curve" are indicated by the intersection of the red dashed line with the black dashed curve. **Panel D**: This is a close up of Panel C. It is clear that the "optimal resistant outbreak curve" is the outbreak curve that is tangent to the κI curve at the point from where it originates from the κI curve.

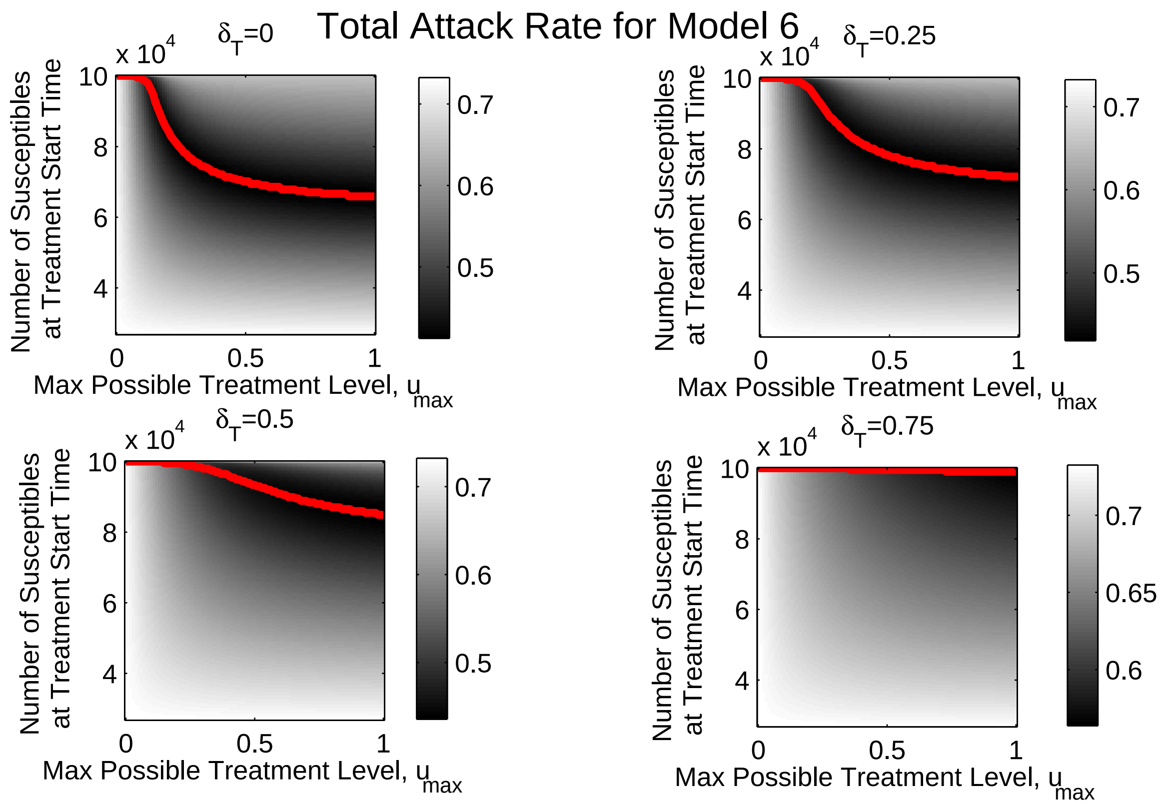

Figure 2.5: Attack rate as a function of maximum treatment level and number of susceptibles at treatment start time for Model (2.6). Each panel corresponds to a different δ_T (if $\delta_T = 0$ then treated individuals do not transmit the disease). Each point on a panel indicates the attack rate that occurs when treatment starts as soon as the number of susceptibles decreases to the value indicated on the vertical axis and continues at the maximum rate (as indicated on the horizontal axis) until the outbreak ends. Notice that in all four panels, as treatment level increases the optimal amount of delay increases. Furthermore, as δ_T increases the optimal amount of delay decreases (when $\delta_T = 0.75$ it is optimal to delay only a very small amount). This trend will be addressed in Section 2.5. Simulations assumed that immediately treating all infected individuals does not immediately end the outbreak. Parameters are $\mu = \frac{1}{3.3}, \frac{\beta}{\mu} = 1.8$, $S_0 = 100000$, $I_0 = 10$, $T(t_0) = 0$, $R(t_0) = 0$, $\nu = 0.002$, $\delta_R = 0.9$.

the number of susceptibles to decrease to $S = S^*$. Determining the optimal treatment strategy when $\delta_T > 0$ remains an open problem but numerical simulations suggest that (i) it may be optimal to delay the start of treatment and (ii) once treatment begins it is best to treat at the maximal rate (i.e., $u = u_{\max}$) and (iii) for any fixed value of $u_{\max}$ the optimal amount of delay is less than the optimal amount of delay when treatment can be administered instantaneously. Figure 2.5 illustrates these observations.

2.5 Discussion

We have demonstrated that by using basic models we can develop simple geometric arguments that help explain some of the dominant effects of treatment. Two key observations are that (i) treatment allows an outbreak to "jump" from one outbreak curve to another and (ii) the shape of any particular outbreak curve is largely determined by the characteristics of the most prevalent strain. Strictly speaking these observations rely on the assumption that (i) treatment can be administered instantaneously and (ii) that individuals who are effectively treated do not transmit the disease. Although these assumptions are not realistic we showed (using numerical simulations) that the general principles derived using these assumptions seem to also hold more generally.

We described how, for our "idealized treatment model", the optimal treatment strategy is the strategy that forces the outbreak to eventually follow the lowest possible outbreak curve. We then discussed what this means in the context of limited resources and the appearance of de novo resistance. More specifically for a single strain model and a limited number of treatment doses we showed that

(i) if there are insufficient resources to treat all of the infected individuals at the start of the outbreak then any treatment strategy that uses all of the available resources is optimal.

Geometrically, this result is a consequence of the fact that all possible outbreak paths are vertical translations of the original outbreak path and therefore, since treatment causes the number of infections to "jump" to a lower value, the total number of treated individuals determines which outbreak path the epidemic eventually follows. We also demonstrated, with numerical simulations, that similar results seem to hold when treatment is not instantaneous and treated individuals still transmit the disease (Figure 2.3). Figure 2.3 suggests that if there are insufficient resources to treat with maximal effort for the entire outbreak then any treatment strategy that uses all the resources is optimal. It is important to emphasize that this is different than saying that the optimal treatment strategy is any strategy that uses all of the available resources. For example, in our idealized model it is best to treat all infected individuals at the start of the outbreak if this is possible. If we have more resources than needed then it is still optimal to treat all infecteds at the start of the outbreak (i.e., we should not wait for the number of infected individuals to increase so that we can use all of our resources). In Chapter 3 it is shown that a similar result holds for Model (2.2) when treated individuals do not transmit the disease (i.e., $\delta_T = 0$). Namely, it is optimal to use all resources only if there are insufficient resources to treat with maximal effort for the entire outbreak. If on the other hand there are sufficient resources to treat with maximal effort for the entire outbreak then this is the best treatment strategy. It is important to emphasize that although Figure 2.3 only considers simple "treatment off"-"treatment on"-"treatment off" strategies, Chapter 3 provides a full proof that considers arbitrary time varying treatment strategies for the case when $\delta_T = 0$. Although it seems reasonable that a similar result holds for the case when treated individuals transmit the disease (and Figure 2.3 supports this hypothesis), further work is required to prove the complete result for the case when $\delta_T > 0$.

For a two strain model (one treatment-sensitive strain and one treatment-resistant strain) we showed that

(i) if treating all infected individuals at the start of the outbreak does not immediately end the outbreak then it is optimal to start treatment when the number of susceptibles is $S^* = \frac{\mu}{\delta_R \beta}(1 + \kappa \frac{1-\delta_R}{1-\kappa})$.

Geometrically this result is a consequence of the fact that the outbreak paths of the resistant strain are not merely vertical translations of the outbreak paths of the sensitive strain. Because the possible resistant outbreak paths originate from the κI curve and the resistant outbreak paths do not cross each other, the lowest possible resistant outbreak path is the one that is initially tangent to the κI curve. Thus the optimal treatment start time is the time when $S = S^*$. Additionally, Figure 2.5 demonstrates that the general rule: "delay and then treat with maximal effort for the remainder of the outbreak" also seems to hold when the rate of treatment is bounded and treated individuals continue to transmit the disease. Figure 2.5 also suggests that the optimal amount of delay for any particular value of $u_{\max}$ is bounded by the optimal amount of delay when treatment can be administered instantaneously (i.e., $u_{\max} \to \infty$). In fact, in Chapter 4 it is shown that for Model (2.6) if $\delta_T = 0$ then the number of susceptibles at the optimal treatment start time is greater than or equal to S^*. Finally, Figure 2.5 suggests that as δ_T increases the upper bound for the optimal amount of delay decreases. For example in panel D ($\delta = 0.75$) the optimal amount of delay (even for large values of $u_{\max}$) is very small. Intuitively this makes sense since for Model (2.6) the sensitive and resistant outbreaks are not temporally separated (recall that in the "idealized treatment model" the sensitive outbreak occurs before treatment and the resistant outbreak occurs after treatment, see Table 2.1). Since there is some temporal overlap of the resistant and sensitive outbreaks, these two outbreaks will compete for susceptibles. If treated individuals are just as infectious as untreated individuals (i.e., as δ_T approaches 1) then it is desirable for the majority of the infections to be resistant (because the resistant strain **is** less transmissible than the untreated sensitive strain). Therefore, it is desirable to start treatment immediately and thus start the resistant outbreak as soon as possible. The competition between the sensitive and resistant strain for susceptibles should help limit the size of the sensitive outbreak.

Chapter 3

Optimal Control of Epidemics with Limited Resources

3.1 Abstract

We extend the existing work on the time-optimal control of the basic SIR epidemic model with mass action contact rate. Previous results have focused on minimizing an objective function that is a linear combination of the cost associated with using control and either the outbreak size or the infectious burden. We instead, provide analytic solutions for the control that minimizes the outbreak size (or infectious burden) under the assumption that there are limited control resources. We provide optimal control policies for an isolation only model, a vaccination only model and a combined isolation-vaccination model (or mixed model). The optimal policies described here contain many interesting features especially when compared to previous analyses. For example, under certain circumstances the optimal isolation only policy is not unique. Furthermore the optimal mixed policy is not simply a combination of the optimal

isolation only policy and the optimal vaccination only policy. The results presented here also highlight a number of areas that warrant further study and emphasize that time-optimal control of the basic SIR model is still not fully understood.

3.2 Introduction

Mathematical analyses can provide valuable information about how best to control infectious disease outbreaks. One issue of real practical concern, and for which mathematical modeling has much to offer, is in determining the optimal distribution of limited resources during an outbreak. Commonly, in preparation for an outbreak, a fixed amount of vaccine and other drugs are stockpiled, in addition to the allocation of a certain amount of funds for other control measures such as isolation and quarantine. Once the epidemic starts, the goal is then to optimally administer these resources given that their supply is limited.

One common aim when administering limited resources is to identify a sub-population within the general population that is best targeted to receive the control resources. Consider for example the Polio outbreak in the Netherlands in 1993-94 [34]. This outbreak was concentrated in an unvaccinated sub-population, and since there was not enough vaccine to treat the entire population, the vaccine was predominantly used in the identified sub-population.

In many cases, however, there is no clearly identifiable sub-population on which to focus control efforts. Furthermore, often the available resources will be insufficient to target all individuals of a sub-population, even if such a group can be identified. As a result, it is critical that the resources are administered in a time-optimal fashion. In other words, if we do not have enough resources to target all appropriate individuals during the entire course of an outbreak, how then should these resources be used over time so as to achieve some specific goal (e.g., minimizing outbreak size)? Such time-optimal control strategies are the focus of this chapter.

Although an analysis of the time-optimal application of outbreak controls is of clear practical value, surprisingly little analysis has been done for the basic SIR model [1–3, 11, 21, 37, 41, 42]. In addition, many of the previous results that have been published have not been fully assimilated by practicing epidemiologists and public health officials. Some of the earliest work in this area is by Abakuks. In [2], Abakuks investigated the optimal control of a simple deterministic SIR model, and determined the isolation strategy that minimizes the total number of infected individuals, balanced against a cost associated with using isolation. He assumed that, at any instant, any number of infectives could be isolated. His results then demonstrated that the optimal control was to either isolate all of the infectives at the beginning of the epidemic, or to never isolate any of them, depending on how heavily the cost of isolation was weighted in the objective function. In [1], Abakuks determined the optimal vaccination strategy for the same model and found that the optimal strategy was to vaccinate N susceptibles at the start of the epidemic, where N depends on the precise form of the objective function. In [3], Abakuks then determined the optimal vaccination strategy for the same model but under the assumption that, at any instant, either all or none of the susceptibles are vaccinated. He found that the optimal control had the same basic form as the optimal control for the isolation model. Shortly after the publication of [2] and [3], Wickwire published [37] and [41] where he studied the same questions but with two notable differences. Firstly, he removed the unrealistic assumption that an arbitrary number of individuals can be isolated or vaccinated instantaneously, and instead modeled these processes under the assumption that some finite rate of isolation or vaccination is possible. Secondly, Wickwire determined the optimal control that minimizes the total infectious burden over an outbreak (plus a cost for using the control) rather than the control that minimizes the total outbreak size. He found that the optimal isolation strategy was to use either maximal control for the entire epidemic or to use no control at all and that the optimal vaccination strategy has at most one switch (switching from maximal vaccination to no vaccination). Wickwire's results therefore, agree qualitatively with Abakuks' results when the control dynamics are changed from acting instantaneously to acting through a rate parameter. Interestingly, however, neither Abakuks nor Wickwire provided results for a combined isolation-vaccination model. In fact, even in [11], which expands Wickwire's results to models with more general contact rates, there is no solution provided for such mixed models.

Our work extends that of [1–3, 37, 41] by examining the kind of resource constraints mentioned earlier. Specifically, the simple SIR model with mass action contact is revisited, and the following question is asked: "Given that there is a limited amount of control resources available during an outbreak, what control policy minimizes the total outbreak size?". Although such SIR models are probably too simplistic for many real outbreaks of interest for which there are limited control resources, we focus on this simple case for two reasons. First, we are able to provide a complete analytical solution for these cases, both for independent isolation and vaccination, as well as for a mixed model that allows both controls simultaneously. Thus, although these models are not intended as definitive solutions to real-world optimal isolation and vaccination problems, they provide a foundation of rigorous mathematical results upon which more complex models can be built, and for which such complete solutions are not possible. Second, even in this very simple epidemiological model, the analysis is not trivial, and the results have some surprising and initially counterintuitive features. For example, the optimal isolation strategy exhibits threshold behaviour, the solution switching from being non-unique to being unique under different parameter values, whereas the optimal vaccination strategy has a consistent basic form. Additionally, in the case of limited resources, the solution for the optimal control strategy of the mixed model is not, in fact, a direct combination of the optimal isolation and vaccination strategies on their own.

The remainder of this chapter is structured as follows. Section 3.3 reviews the basic model used in all further analysis and gives a precise mathematical statement of the problem. Section 3.4 describes the optimal isolation, optimal vaccination and optimal mixed polices under limited resources. The next three sections then present the proofs for the optimal isolation, optimal vaccination and optimal mixed polices, respectively. Lastly, Section 3.8 discusses some of the salient features of the results.

3.3 Model and Setup

The following standard deterministic SIR model is used throughout this chapter [7]:

$$\begin{aligned} \dot{S} &= -\beta S I - u_v S, \\ \dot{I} &= \beta S I - (\mu + u_i) I, \end{aligned} \tag{3.1}$$

where S and I are the numbers of susceptible and infected hosts (a dot indicating time derivatives), β is the transmission rate, μ is the per capita loss rate of infected individuals through both mortality and recovery, and u_v and u_i are the per capita rates of vaccination and isolation respectively. Note that, by definition, vaccination has a direct effect only on susceptible individuals whereas isolation has a direct effect only on infected individuals. If $u_v \equiv 0$ then (3.1) describes the standard isolation model, if $u_i \equiv 0$ then (3.1) describes the standard vaccination model, and if neither u_v nor u_i are the zero function then (3.1) describes the isolation-vaccination model (or mixed model). Note that we do not explicitly track the dynamics of the recovered sub-population.

The main reason we have chosen to analyze model (3.1) is that it allows us to extend the work of [1–3, 37, 41] to the case when there are limited resources. This being said, it is worth mentioning some of the major assumptions implicit in model (3.1). The first major assumption is that disease transmission is well modelled using mass action incidence. Although it is perhaps more common to use standard incidence when modeling human disease, there are important studies that use mass action incidence (for example a recent study for influenza [28]). We have therefore decided to use mass action incidence because it is consistent with the models used in [1–3, 37, 41] and because there is a precedent for it in current literature. The second major assumption is that isolation and vaccination are completely effective. In other words, isolated individuals cannot spread the infection and vaccinated individuals cannot contract the infection. Although we have not undertaken the detailed analysis required to fully understand the implications of these assumptions, we have performed some preliminary numerical analysis that suggests that these assumptions do not significantly alter the main conclusions of our work (see section 3.10).

Before proceeding further, it is worth discussing the bounds on the control variables, u_i and u_v in (3.1). The variables u_i and u_v are the per-capita rates of isolation and vaccination respectively, and therefore in practice each of these will be related to the isolation and vaccination effort that is employed. For example, the level of u_i that can be attained at any given time will, to some extent, depend upon the public health infrastructure (e.g., number of public health officers available). At the same time, however, the characteristics of the disease will impose an upper limit on the possible values of u_i. For example, if some infections are asymptomatic (or if there is an asymptomatic phase), then even if we employed an infinite isolation effort, u_i would remain bounded. This simply reflects the fact that, regardless of how much we strive to increase the rate of isolation or vaccination, it is usually unavoidable that the targeted individuals will spend some nonzero amount of time circulating in the population before they are recruited (i.e., before they are isolated of vaccinated). Therefore, we model this by specifying an upper bound for both control variables, and suppose that this upper bound is disease-specific. The presence of these upper bounds on the control variables is distinct from our assumption (to be formalized shortly) that control resources are limited. The upper bounds reflect unavoidable constraints on the rate at which we can isolate or vaccinate, whereas the assumption of limited resources reflects constraints on the total number of people we can isolate or vaccinate.

Remark 3.1. (i) Let $z(t)$ denote the total number of vaccinated susceptibles at time t and let $w(t)$ denote the total number of isolated individuals at time t. The actual values of S, I, z and w will depend on the specific choice of controls

u_i and u_v. When necessary, in order to make this dependence explicit, $S_{[u_i,u_v]}$, $I_{[u_i,u_v]}$, $z_{[u_i,u_v]}$ and $w_{[u_i,u_v]}$ will be used to denote the states satisfying (3.1) with the particular choice of controls, u_i and u_v. Furthermore, if $u_i \equiv 0$ ($u_v \equiv 0$) then $S_{[u_v]}$ ($S_{[u_i]}$) will be used instead of $S_{[0,u_v]}$ ($S_{[u_i,0]}$).

(ii) Let $c \in \mathbb{R}$. Using a slight abuse of notation, if $u_i(t) = c$ ($u_v(t) = c$) for all time then u_i (u_v) will be denoted by $u_i = c$ ($u_v = c$).

The general question addressed in this chapter can be phrased mathematically as follows:

General Problem

Fix $w_{\max} \geq 0$ and $z_{\max} \geq 0$. Determine the control for model (3.1) that minimizes

$$\int_{t_0}^{T} \beta S_{[u_i,u_v]} I_{[u_i,u_v]} \mathrm{d}t, \tag{3.2}$$

subject to $S(t_0) = S_0$, $I(t_0) = I_0$, $T = \inf\{t \mid I_{[u_i,u_v]}(t) = 0.5\}$, $(u_i(t), u_v(t)) \in [0, u_{m_i}] \times [0, u_{m_v}]$ for all $t \in [0, T]$, and subject to the resource constraints

$$\int_{t_0}^{T} u_i I_{[u_i,u_v]} \mathrm{d}t \leq w_{\max}, \tag{3.3}$$

and

$$\int_{t_0}^{T} u_v S_{[u_i,u_v]} \mathrm{d}t \leq z_{\max}. \tag{3.4}$$

All of the results presented here hold for any $u_{m_i} \in (0, \infty)$ and $u_{m_v} \in (0, \infty)$ but to simplify notation it is assumed that $u_{m_i} = u_{m_v} = u_{\max}$. Expression (3.2) represents the total outbreak size over the course of the epidemic, and expressions (3.3) and

(3.4) reflect the assumption that there is a maximum total amount of isolation and vaccination, respectively, that can be used during the course of the epidemic.

The method of how T was chosen in Problem 3.3 requires some further explanation. The end of the epidemic in models such as system (3.1) is not well defined, because the number of infected individuals approaches zero asymptotically as time approaches infinity. As a result, it is not immediately clear how best to choose T. The three most natural choices are

(i) $T = \infty$,

(ii) $T = T_{\max}$, where $T_{\max}$ is a sufficiently large constant, and

(iii) $T = \inf\{t \mid I(t) = I_{\min}\}$, where $I_{\min}$ is some constant chosen to indicate the end of the epidemic.

A potential problem with choosing methods (i) and (ii) is that the structure of the optimal control can be affected by the dynamics of the system at very large values of t in undesirable ways. For example, suppose u_i is nonzero and then at some time t_1, $I(t_1) < 1$ and $w(t_1) = w_{\max}$. Then because constraint (3.3) is now active, $u_i(t) = 0$ for all $t > t_1$. If $\dot{I}(t_1^+) > 0$, however, then there will be another peak in infectives before the epidemic settles down permanently. This sort of behaviour is undesirable because the final peak in infectives is caused by a fractional number of infectives. Furthermore, an additional complication of method (ii) is that it is not clear how large T must be in order to coincide with a reasonable interpretation of the end of the epidemic. Method (iii) avoids both of these difficulties because it terminates the epidemic as soon as $I(t) = I_{\min}$. In the general problem statement above, $I_{\min}$ was chosen to be 0.5 but any value smaller than 1 would work just as well.

In order to make the solution to the above General Problem more transparent, we break it down into three separate problems:

Problem 3.1. (Isolation Only) Solve General Problem with $z_{\max} = 0$.

Problem 3.2. (Vaccination Only) Solve General Problem with $w_{\max} = 0$.

Problem 3.3. (Mixed Model) Solve General Problem with $z_{\max} > 0$ and $w_{\max} > 0$.

The primary method applied to solve these problems is Pontryagin's Maximum principle (PMP). (See [13, 26, 35] for details of PMP.) A simple application of Filippov's theorem, [4], shows that optimal controls exist for Problems 3.1, 3.2 and 3.3. The proof of the existence of an optimal control is detailed in Appendix A.

Finally, it is worthwhile to point out that, although the General Problem minimizes the total number of infectives, $\int_{t_0}^{T} \beta IS\mathrm{d}t$, this is equivalent to minimizing the total infectious burden $\int_{t_0}^{T} I\mathrm{d}t$. For Problem 3.1, $\int_{t_0}^{T} \beta IS\mathrm{d}t = S_0 - S(T)$ and $\int_{t_0}^{T} I\mathrm{d}t = -\frac{1}{\beta}\ln(\frac{S(T)}{S_0})$, and therefore both quantities are minimized by maximizing $S(T)$. For Problem 3.2, notice that $\int_{t_0}^{T} \beta IS\mathrm{d}t = I(T) - I_0 + \mu \int_{t_0}^{T} I\mathrm{d}t$. Since $I(T)$ and I_0 are fixed, minimizing the total number of infectives and minimizing the total infectious burden are equivalent. For Problem 3.3 it is also the case that both cost functions lead to the same optimal control. Although this is perhaps less obvious, it can be proven by using PMP to solve Problem 3.3 with (3.2) replaced by $\int_{t_0}^{T} I\mathrm{d}t$. The procedure to do this is very similar to the solution of Problem 3.3 and so is not detailed further.

3.4 Optimal Policies

The following three theorems provide solutions to Problems 1-3.

Theorem 3.1. (Optimal Isolation Policy) If $w_{[u_{\max}]}(T) \leq w_{\max}$, then the optimal isolation policy for Problem 3.1 is $u_i^* \equiv u_{\max}$. If $w_{[u_{\max}]}(T) > w_{\max}$, then the optimal isolation policy u_i^* is any control u_i such that $w_{[u_i]}(T) = w_{\max}$.

A key feature affecting the optimal isolation policy in Theorem 3.1 is whether or not the resources are sufficient to maintain a maximal isolation rate for the entire duration of the outbreak (i.e., $w_{[u_{\max}]}(T) \leq w_{\max}$ versus $w_{[u_{\max}]}(T) > w_{\max}$). Whether or not this is the case will depend on several factors. First, it is clear that the disease-specific transmissibility, β, as well as the per-capita removal rate, μ, will affect whether or not the resources are sufficient. Less obvious, however, is that the maximum possible isolation rate, $u_{\max}$, will also play a role. For example, if $u_{\max}$ is relatively large, then it might be possible to reduce the size of the outbreak substantially, thereby requiring a smaller overall total amount of resources. On the other hand, if $u_{\max}$ is relatively small, then a large outbreak might be unavoidable, meaning that a much larger total pool of resources would be required to maintain a maximal isolation effort for the entire epidemic. Thus, the way the disease in question affects the value of $u_{\max}$ (as well as how the existing public health infrastructure affects $u_{\max}$) will play an important role in deciding which of the two outcomes in Theorem 3.1 occur.

Theorem 3.1 says that if there are sufficient isolation resources, then the optimal strategy is to isolate infectives with maximal effort for the entire epidemic. If there are not sufficient resources to do this, then the optimal strategy is any strategy that uses all of the available isolation resources. So, for example, it is optimal to start isolating with maximal effort at the start of the epidemic and to continue isolating with maximal effort until all of the resources have run out, but this is only one of many possible optimal strategies.

This non-uniqueness of the optimal solution deserves further consideration as it gives policy makers much freedom in choosing how to isolate individuals. One complication is that exercising this freedom requires knowing a priori that the available resources are insufficient. Data from previous epidemics as well as case reports of the current epidemic can be used to estimate the required resources. Alternatively, in very serious situations, it might be quite obvious that the available resources are insufficient. At the same time, however, this flexibility means that policy makers also have the freedom to optimize other cost functions in addition to (3.2). For example, if one wanted to minimize the average number of infected individuals per unit time over the entire outbreak, it is shown in Section 3.9 that the optimal policy is to isolate with maximal effort as long as possible. Given that this is also an optimal strategy for the cost function (3.2), it will simultaneously optimize both criteria.

Theorem 3.2. (Optimal Vaccination Policy) There exists a $\tau \in [0, T]$ such that the optimal vaccination policy for Problem 3.2 is

$$u_v^*(t) = \begin{cases} u_{\max} & t \in [0, \tau) \\ 0 & t \in [\tau, T], \end{cases} \tag{3.5}$$

where $\int_{t_0}^{\tau} u_{\max} S_{[u_{\max}]} \mathrm{d}t = z_{\max}$ if $\tau < T$. That is, the optimal vaccination policy is to vaccinate with maximal effort until either all of the resources are used up or the epidemic is over.

Theorem 3.2 reveals that the optimal vaccination policy is unique, with the strategy being to employ maximal vaccination from the beginning of the outbreak, for as long as possible. An interesting consequence of the above two theorems is that, under certain circumstances, the optimal isolation policy is less sensitive to initial conditions than the optimal vaccination policy. For example, if there are insufficient resources to isolate maximally for the entire epidemic then there is no penalty for delaying an isolation program provided the maximal possible number of individuals are still isolated. This is not, however, true for vaccination.

Theorem 3.3. (Optimal Mixed Policy) There exists a $\tau \in [0, T]$ such that the optimal mixed isolation and vaccination strategy for Problem 3.3 has one of the following forms:

(i)

$$(u_i^*(t), u_v^*(t)) = \begin{cases} (u_{\max}, u_{\max}) & t \in [0, \tau] \\ (0, u_v^*(t)) & t \in (\tau, T], \end{cases} \tag{3.6}$$

where $\int_{t_0}^{\tau} u_{\max} I_{[u_{\max}, u_{\max}]} \mathrm{d}t = w_{\max}$ if $\tau < T$, or

(ii)

$$(u_i^*(t), u_v^*(t)) = \begin{cases} (u_{\max}, u_{\max}) & t \in [0, \tau] \\ (u_i^*(t), 0) & t \in (\tau, T], \end{cases} \tag{3.7}$$

where $\int_{t_0}^{\tau} u_{\max} S_{[u_{\max}, u_{\max}]} \mathrm{d}t = z_{\max}$ if $\tau < T$.

In Equation (3.6), if $t > \tau$ then $u_v^*(t)$ represents the optimal control for the corresponding vaccination only model. That is, for $t > \tau$, $u_v^*(t)$ is the solution to Problem 3.2 with $\tilde{z}_{\max} = z_{\max} - \int_{t_0}^{\tau} u_{\max} S_{[u_{\max}, u_{\max}]} \mathrm{d}t$ and initial conditions $\tilde{t}_0 = \tau$, $\tilde{S}_0 = S_{[u_{\max}, u_{\max}]}(\tau)$ and $\tilde{I}_0 = I_{[u_{\max}, u_{\max}]}(\tau)$. In Equation (3.7), $u_i^*(t)$ for $t > \tau$ is defined similarly.

Theorem 3.3 says that the optimal isolation-vaccination strategy is to isolate and vaccinate with maximal effort until either the vaccination or isolation resources run out. If the isolation resources run out first, then it is optimal to continue with the optimal vaccination strategy. If the vaccination resources run out first, then it is optimal to continue with the optimal isolation strategy. The optimal mixed policy is particularly interesting because it is not simply a combination of the optimal isolation and the optimal vaccination policies. Consider the case where there are insufficient isolation resources to isolate with maximal effort for the entire epidemic. Then, unlike the optimal isolation policy, u_i^* is no longer given by any control that uses all of the isolation resources. In fact the isolation effort must be maximal at the start of the epidemic and must remain maximal until either the isolation or the vaccination resources run out. If the vaccination resources run out first, then the isolation policy for the remainder of the outbreak can then be anything that uses the remaining isolation resources.

3.5 Proof for Optimal Isolation Policy

The isolation model with limited resources is described by the following set of equations:

$$\begin{aligned} \dot{S} &= -\beta S I, && (3.8) \\ \dot{I} &= \beta S I - (\mu + u_i) I, && (3.9) \\ \dot{w} &= u_i I. \end{aligned}$$

Phrasing Problem 3.1 as a maximization problem and recalling that a free terminal time implies that $H = 0$, PMP provides the following relations:

The Hamiltonian is

$$\begin{aligned} H(t) &= -\lambda_0\beta SI - \lambda_S\beta SI + \lambda_I\beta SI - \lambda_I(\mu + u_i)I + \lambda_w u_i I \\ &= -\dot{\lambda}_I I = -\dot{\lambda}_S S - \lambda_I\mu + (\lambda_w - \lambda_I)u_i I = 0, \end{aligned} \tag{3.10}$$

where the adjoint variables satisfy,

$$\dot{\lambda}_S = -(\lambda_I - \lambda_0 - \lambda_S)\beta I, \tag{3.11}$$

$$\dot{\lambda}_I = -(\lambda_I - \lambda_0 - \lambda_S)\beta S - (\lambda_w - \lambda_I)u_i + \lambda_I\mu, \tag{3.12}$$

$$\dot{\lambda}_w = 0.$$

The transversality conditions are $(\lambda_0, \lambda_S(T), \lambda_I(T), \lambda_w) = (\lambda_0, 0, \lambda_I(T), q)$, where $q \leq 0$. The optimal control therefore satisfies

$$u_i^* = \begin{cases} u_{\max} & \lambda_w > \lambda_I \\ ? & \lambda_w = \lambda_I \\ 0 & \lambda_w < \lambda_I. \end{cases}$$

From Equation (3.10) we have $\dot{\lambda}_I = 0$, and therefore the optimal control is either $u_i^* \equiv 0$, $u_i^* \equiv u_{\max}$ or u_i^* is singular.

The, solution to Problem 3.1 uses the following three observations:

(i) Without the constraint (3.3) the solution to Problem 3.1 is $u_i^* \equiv u_{\max}$. In the sequel this solution will be called the unconstrained optimal control. The proof of this result is given in Section 3.9.

(ii) The total number of isolated individuals can be written as:

$$\int_{t_0}^{T} u_i I_{[u_i]} \mathrm{d}t = S_0 - S_{[u_i]}(T) + I_0 - I_{[u_i]}(T) + \frac{\mu}{\beta}\ln\left(\frac{S_{[u_i]}(T)}{S_0}\right). \tag{3.13}$$

Equation (3.13) shows that the constraint value, $w_{[u_i]}(T) = \int_{t_0}^{T} u_i I_{[u_i]} \mathrm{d}t$, depends only on $S_{[u_i]}(T)$ (see Section 3.9 for details).

(iii) The cost function can be rewritten as $\int_{t_0}^{T} \beta I_{[u_i]} S_{[u_i]} \mathrm{d}t = S_0 - S_{[u_i]}(T)$ and therefore minimizing the cost function is equivalent to maximizing $S_{[u_i]}(T)$.

Clearly, if $w_{[u_{\max}]}(T) \leq w_{\max}$ then the optimal control is the unconstrained optimal control $u_i^* \equiv u_{\max}$. To determine the optimal control when $w_{[u_{\max}]}(T) > w_{\max}$ first notice that Equation (3.13) can be rewritten as,

$$\frac{\mu}{\beta}\ln(S_{[u_i]}(T)) - S_{[u_i]}(T) = I_{[u_i]}(T) - I_0 - S_0 + \frac{\mu}{\beta}\ln(S_0) + w_{[u_i]}(T).$$

There are two possible scenarios:

(i) Suppose $w_{[u_{\max}]}(T) > w_{\max}$ and $S_{[u_{\max}]}(T) < \frac{\mu}{\beta}$. Then $S_{[u_i]}(T)$ is an increasing function of $w_{[u_i]}(T)$ for any $w_{[u_i]}(T) \leq w_{\max} < w_{[u_{max}]}(T)$. Therefore, u_i^* is any control that uses all of the available resources.

(ii) Suppose $w_{[u_{\max}]}(T) > w_{\max}$ and $S_{[u_{\max}]}(T) > \frac{\mu}{\beta}$. Since $f(S) = \frac{\mu}{\beta}\ln(S) - S$ is a convex down function with maximum $S = \frac{\mu}{\beta}$, for any u_i satisfying $w_{[u_i]}(T) < w_{\max}$ it must be that $S_{[u_i]}(T) < \frac{\mu}{\beta}$. Then $S_{[u_i]}(T)$ is an increasing function of $w_{[u_i]}(T)$ for any $w_{[u_i]}(T) < w_{\max}$. Therefore, u_i^* is any control that uses all of the available resources.

This concludes the proof of Theorem 3.1.

3.6 Proof for Optimal Vaccination Policy

The vaccination model with limited resources is described by the following set of equations:

$$\begin{aligned}\dot{S} &= -\beta SI - u_v S,\\ \dot{I} &= \beta SI - \mu I,\\ \dot{z} &= u_v S.\end{aligned}$$

Phrasing Problem 3.2 as a maximization problem, PMP provides the following relations:

The Hamiltonian is

$$\begin{aligned}H(t) &= -\lambda_0\beta SI - \lambda_S\beta SI - \lambda_S u_v S + \lambda_I\beta SI - \lambda_I\mu I + \lambda_z u_v S,\\ &= -\dot{\lambda}_I I + (\lambda_z - \lambda_S)u_v S = -\dot{\lambda}_S S - \lambda_I\mu I = 0, \qquad (3.14)\end{aligned}$$

where the adjoint variables satisfy,

$$\begin{aligned}\dot{\lambda}_S &= -(\lambda_I - \lambda_0 - \lambda_S)\beta I - (\lambda_z - \lambda_S)u_v, \qquad (3.15)\\ \dot{\lambda}_I &= -(\lambda_I - \lambda_0 - \lambda_S)\beta S + \lambda_I\mu, \qquad (3.16)\\ \dot{\lambda}_z &= 0.\end{aligned}$$

The transversality conditions are $(\lambda_0, \lambda_S(T), \lambda_I(T), \lambda_z) = (\lambda_0, 0, \lambda_I(T), p)$, where $p \leq$

0. The optimal control therefore satisfies

$$u_v^* = \begin{cases} u_{\max} & \lambda_z > \lambda_S \\ ? & \lambda_z = \lambda_S \\ 0 & \lambda_z < \lambda_S. \end{cases} \tag{3.17}$$

The first step is to show that the optimal control is purely bang-bang (that is, has no singular components). If $\lambda_z = \lambda_S$ on some interval J then $\dot{\lambda}_S = 0$ on J. Equation (3.15) then gives,

$$\begin{aligned} 0 &= -(-\lambda_0 - \lambda_S + \lambda_I)\beta I - (\lambda_z - \lambda_S)u_v, \\ 0 &= -(-\lambda_0 - \lambda_S + \lambda_I)\beta I, \\ 0 &= (-\lambda_0 - \lambda_S + \lambda_I), \\ \lambda_I &= \lambda_0 + \lambda_S, \end{aligned} \tag{3.18}$$

and therefore $\dot{\lambda}_I = 0$ on J. If $\dot{\lambda}_I = 0$, then by Equations (3.16) and (3.18) it must be that $\lambda_I = 0$. Therefore $\lambda_S = -\lambda_0$ and the only nonzero choice for the adjoint variables on J is $(\lambda_0, \lambda_S, \lambda_I, \lambda_z) = (1, -1, 0, -1)$. Further, by Equations 3.15 and 3.16, once u_v^* becomes singular then it must remain singular (that is, $T \in J$). Now since $T \in J$, $(\lambda_0, \lambda_S, \lambda_I, \lambda_z) = (1, -1, 0, -1)$ must satisfy the transversality condition $\lambda_S(T) = 0$. Since the transversality condition is not satisfied, the optimal control must be purely bang-bang.

Next, we consider the times at which the optimal control switches between 0 and

$u_{\max}$. Suppose the optimal control switches at times t_{s_i}. Then

$$H(t_{s_i}) = -\dot{\lambda}_I(t_{s_i})I(t_{s_i}) = -\dot{\lambda}_S(t_{s_i})S(t_{s_i}) - \lambda_I(t_{s_i})\mu I(t_{s_i}) = 0. \tag{3.19}$$

Therefore, substituting $\dot{\lambda}_I(t_{s_i}) = 0$ into Equation (3.16) gives,

$$(\lambda_S(t_{s_i}) + \lambda_0)\beta S(t_{s_i}) \quad = \quad \lambda_I(t_{s_i})(\beta S(t_{s_i}) - \mu). \tag{3.20}$$

The rest of the proof relies on the following relations which are obtained from Equation (3.17) and Equation (3.19):

$$\begin{array}{lllll} \lambda_I(t_{s_i}) > 0 & \Rightarrow & \dot{\lambda}_S(t_{s_i}) < 0 & \Rightarrow & (0 \rightarrow u_{\max}) \\ \lambda_I(t_{s_i}) < 0 & \Rightarrow & \dot{\lambda}_S(t_{s_i}) > 0 & \Rightarrow & (u_{\max} \rightarrow 0) \\ \lambda_I(t_{s_i}) = 0 & \Rightarrow & \dot{\lambda}_S(t_{s_i}) = 0 & \Rightarrow & \text{no switch occurs.} \end{array} \tag{3.21}$$

Remark 3.2. In (3.21), the notation $(a \rightarrow b)$ means that u_v^* switches from the value a to the value b at $t = t_{s_i}$. In the sequel, the notation:

$$\begin{array}{ccccc} a & \rightarrow & b & \rightarrow & b \\ c & \rightarrow & d & \rightarrow & e \end{array}$$

describes two possible forms for u_v^*. In the first form, u_v^* has the value a for some time and then switches to the value b for the remainder of the time. In the second form, u_v^* has the value c for some time and then switches to the value d for some time and then switches to the value e for the remainder of the time.

Now since $\lambda_S(t_{s_i}) = \lambda_z =$ constant, $\lambda_S(t_{s_i}) + \lambda_0$ is either always positive, always negative or always zero.

Suppose $\lambda_S(t_{s_i}) + \lambda_0 = 0$. Then by (3.20), either $\lambda_I(t_{s_i}) = 0$ or $S(t_{s_i}) = \frac{\mu}{\beta}$. By (3.21), if $\lambda_I(t_{s_i}) = 0$ then no switch occurs. Alternatively if $S(t_{s_i}) = \frac{\mu}{\beta}$ then the optimal control has only one switch and this switch occurs when I is maximal. The possible optimal controls are:

$$\begin{array}{lcc} u_{\max} & \to & u_{\max} \\ u_{\max} & \to & 0 \\ 0 & \to & u_{\max} \\ 0 & \to & 0 \end{array} \tag{3.22}$$

Next consider the case when $\lambda_S(t_{s_i})+\lambda_0 > 0$. By Equation 3.20, if $\lambda_S(t_{s_i})+\lambda_0 > 0$ then either

(i) $\lambda_I(t_{s_i}) > 0$ and $S(t_{s_i}) > \frac{\mu}{\beta}$ or

(ii) $\lambda_I(t_{s_i}) < 0$ and $S(t_{s_i}) < \frac{\mu}{\beta}$.

Suppose that at t_{s_1} possibility (i) holds. Then by (3.21), u_v^* switches from 0 to $u_{\max}$ at t_{s_1}. If there is to be another switch (say at $t = t_{s_2}$) then it must be from $u_{\max}$ to 0. This implies, by (3.21), that (ii) must hold at t_{s_2}. If there is to be a third switch (say at $t = t_{s_3}$) then it must be from 0 to $u_{\max}$ and therefore, by (3.21), (i) must hold at t_{s_3}. But this is not possible because S is a monotonically decreasing function of time. Similar reasoning applies to the case when possibility (ii) holds at t_{s_1}. Therefore, if $\lambda_S(t_{s_i}) + \lambda_0 > 0$ the optimal control can have at most two switches. The possible forms for the optimal control are:

$$\begin{array}{lllll}
u_{\max} & \rightarrow & u_{\max} & \rightarrow & u_{\max} \\
u_{\max} & \rightarrow & 0 & \rightarrow & 0 \\
0 & \rightarrow & u_{\max} & \rightarrow & 0 \\
0 & \rightarrow & u_{\max} & \rightarrow & u_{\max} \\
0 & \rightarrow & 0 & \rightarrow & 0
\end{array} \tag{3.23}$$

Notice however, that the control form $0 \rightarrow u_{\max} \rightarrow 0$ is not optimal. To see this, recall that the Hamiltonian can be written as $H = -\dot{\lambda}_I I + (\lambda_z - \lambda_S) u_v S$ and therefore $\dot{\lambda}_I(t) \geq 0$ for all $t \in [0, T]$. Also, by (3.21), $\lambda_I(t_{s_1}) > 0$ and $\lambda_I(t_{s_2}) < 0$, but this is not possible since λ_I is monotonically increasing.

Next consider the case when $\lambda_S(t_{s_i}) + \lambda_0 < 0$. Applying similar arguments to those that were used for the case when $\lambda_S(t_{s_i}) + \lambda_0 > 0$, the possible forms for the optimal control are:

$$\begin{array}{lllll}
u_{\max} & \rightarrow & u_{\max} & \rightarrow & u_{\max} \\
u_{\max} & \rightarrow & 0 & \rightarrow & 0 \\
0 & \rightarrow & u_{\max} & \rightarrow & u_{\max} \\
u_{\max} & \rightarrow & 0 & \rightarrow & u_{\max} \\
0 & \rightarrow & 0 & \rightarrow & 0
\end{array} \tag{3.24}$$

Notice however, that the control form, $u_{\max} \rightarrow 0 \rightarrow u_{\max}$ is not optimal because at the first switch $\lambda_I(t_{s_1}) < 0$ and at the second switch $\lambda_I(t_{s_2}) > 0$. This implies that λ_I is non-constant while $u_v^* = 0$ which contradicts Equation (3.14).

Combining the control forms in (3.22), (3.23) and (3.24) gives the following candidates for the optimal control:

$$\begin{array}{lllll} u_{\max} & \rightarrow & u_{\max} & \rightarrow & u_{\max} \\ u_{\max} & \rightarrow & 0 & \rightarrow & 0 \\ 0 & \rightarrow & u_{\max} & \rightarrow & u_{\max} \\ 0 & \rightarrow & 0 & \rightarrow & 0 \end{array} \tag{3.25}$$

The next step is to show that the control forms, $0 \rightarrow u_{\max} \rightarrow u_{\max}$ and $0 \rightarrow 0 \rightarrow 0$, are not optimal. Recall that $\dot{\lambda}_I \geq 0$. Therefore, since $\lambda_I(t_{s_i}) > 0$, it must be that $\lambda_I(t) > 0$ for all $t \in [t_{s_1}, T]$. Therefore, by Equation (3.14), $\dot{\lambda}_S(t) < 0$ for all $t \in [t_{s_1}, T]$. Since $\lambda_S(T) = 0$ this implies that $\lambda_S(t) > 0$ for all $t \in [t_{s_1}, T)$ and therefore $\lambda_S(t_{s_1}) = \lambda_z > 0$ but this is a contradiction since $\lambda_z \leq 0$. The proof that $u_v \equiv 0$ is not optimal is given in Section 3.9.

The discussion above shows that the optimal control is given by,

$$u^*(t) = \begin{cases} u_{\max} & t \in [0, t_{s_1}) \\ 0 & t \in [t_{s_1}, T]. \end{cases} \tag{3.26}$$

To prove the assertion about t_{s_1}, suppose that $u \equiv u_{\max}$ is not optimal. Then, since the optimal control has the form $u_{\max} \rightarrow 0$, by (3.21), $\lambda_I(t_{s_1}) < 0$. This implies, since λ_I is monotonically increasing, that $\lambda_I(t) < 0$ for all $t \in [0, t_{s_1}]$. This implies, since $\dot{\lambda}_I(t) = 0$ for all $t \in [t_{s_1}, T]$, that $\lambda_I(t) < 0$ for all $t \in [0, T]$. By Equation (3.14), this shows that $\dot{\lambda}_S > 0$ for all $t \in [0, T)$ and therefore $\lambda_S(t) < 0$ for all $t \in [0, T)$ (since $\lambda_S(T) = 0$). Therefore, if $t_{s_1} < T$ then $\lambda_S(t_{s_1}) = \lambda_z < 0$. But $\lambda_z < 0$ only if $z_{\max} = z(T)$ and therefore all of the resources must be used up.

3.7 Proof for Optimal Mixed Policy

The mixed model with limited resources is described by the following set of equations:

$$\begin{aligned}
\dot{S} &= -\beta SI - u_v S, \\
\dot{I} &= \beta SI - (\mu + u_i)I, \\
\dot{w} &= u_i I, \\
\dot{z} &= u_v S.
\end{aligned}$$

Phrasing Problem 3.3 as a maximization problem, PMP provides the following relations:

The Hamiltonian is

$$\begin{aligned}
H(t) &= -\lambda_0 \beta SI - \lambda_S \beta SI - \lambda_S u_v S + \lambda_I \beta SI - \lambda_I(\mu + u_i)I + \lambda_w u_i I + \lambda_z u_v S, \\
&= -\dot{\lambda}_I I + (\lambda_z - \lambda_S)u_v S = -\dot{\lambda}_S S + (\lambda_w - \lambda_I)u_i I - \lambda_I \mu I = 0,
\end{aligned} \tag{3.27}$$

where the adjoint variables satisfy,

$$\begin{aligned}
\dot{\lambda}_S &= -(\lambda_I - \lambda_0 - \lambda_S)\beta I - (\lambda_z - \lambda_S)u_v, \\
\dot{\lambda}_I &= -(\lambda_I - \lambda_0 - \lambda_S)\beta S - (\lambda_w - \lambda_I)u_i + \lambda_I \mu, \\
\dot{\lambda}_z &= 0, \\
\dot{\lambda}_w &= 0.
\end{aligned} \tag{3.28}$$

The transversality conditions are $(\lambda_0, \lambda_S(T), \lambda_I(T), \lambda_z, \lambda_w) = (\lambda_0, 0, \lambda_I(T), p, q)$, where

p, $q \leq 0$. The optimal control therefore satisfies

$$u_v^* = \begin{cases} u_{\max} & \lambda_z > \lambda_S \\ ? & \lambda_z = \lambda_S \\ 0 & \lambda_z < \lambda_S, \end{cases}$$

and

$$u_i^* = \begin{cases} u_{\max} & \lambda_w > \lambda_I \\ ? & \lambda_w = \lambda_I \\ 0 & \lambda_w < \lambda_I. \end{cases}$$

Remark 3.3. (i) In the sequel u_i and u_v will denote the optimal controls (instead of u_i^* and u_v^*). The notation u_v^* and u_i^* will denote the optimal controls for the vaccination only model and isolation only model (that is, solutions to Problems 3.2 and 3.1 respectively with appropriate choices of $\tilde{t}_0$, $\tilde{z}_{\max}$, $\tilde{w}_{\max}$, $\tilde{\mu}$, $\tilde{I}_0$ and $\tilde{S}_0$).

(ii) In the sequel, the notation:

$$\begin{array}{rccccccc} u_i: & a & \rightarrow & b & \rightarrow & b & \rightarrow \\ u_v: & c & \rightarrow & d & \rightarrow & e & \rightarrow \\ t: & 0 & \rightarrow & t_{s_1} & \rightarrow & t_{s_2} & \rightarrow \end{array}$$

means that there are two possible switch times, t_{s_1} and t_{s_2}, and u_i switches from a to b at t_{s_1} and does not switch at time t_{s_2}. Similarly, u_v switches from c to d at time t_{s_1} and from d to e at time t_{s_2}.

The solution to Problem 3.3 uses the following claims (the proofs of which can be found in Section 3.9).

Claim 3.3.1. *If u_v is singular on some interval J then u_i is not singular on J and $u_i = 0$ on J.*

Claim 3.3.2. *If u_i is singular on an interval J, then $u_v = 0$ on J.*

Claim 3.3.3. *When u_i is singular, denote it by u_s. If u_i has a singular component, then u_i has one of the following forms:*

$$\begin{array}{lcl} u_{\max} & \rightarrow & u_{\mathrm{s}} \\ u_{\mathrm{s}} & \rightarrow & u_{\mathrm{s}}. \end{array}$$

Claim 3.3.4. *If there exists a $t_s \geq 0$ such that u_i is constant for all $t \in (t_s, T]$ then $u_v(t) = u_v^*(t)$ for all $t \in (t_s, T]$.*

Claim 3.3.5. *If there exists a $t_s \geq 0$ such that $u_v(t) = 0$ for all $t \in (t_s, T]$ then $u_i(t) = u_i^*(t)$ for all $t \in (t_s, T]$.*

Claim 3.3.6. *Let J be a maximal interval such that*

(i) $u_i = u_{\max}$ on J and

(ii) u_i is not singular on J.

If $T \notin J$ then $u_v = u_{\max}$ on J.

Claims 3.3.1-3.3.6 and the fact that $\dot{\lambda}_I \geq 0$, imply that the optimal control must have one of the following forms:

$$\begin{array}{rcl} u_i: & u_s & \rightarrow \\ u_v: & 0 & \rightarrow \\ t: & 0 & \rightarrow, \end{array} \tag{3.29}$$

$$\begin{aligned} u_i:&\quad 0 \quad \rightarrow \\ u_v:&\quad u_v^* \quad \rightarrow \\ t:&\quad 0 \quad \rightarrow, \end{aligned} \tag{3.30}$$

$$\begin{aligned} u_i:&\quad u_{\max} \quad \rightarrow \\ u_v:&\quad u_v^* \quad \rightarrow \\ t:&\quad 0 \quad \rightarrow, \end{aligned} \tag{3.31}$$

$$\begin{array}{rccccc} u_i: & u_{\max} & \rightarrow & u_s & \rightarrow \\ u_v: & u_{\max} & \rightarrow & 0 & \rightarrow \\ t: & 0 & \rightarrow & t_{s_1} & \rightarrow, \end{array} \tag{3.32}$$

$$\begin{array}{rccccc} u_i: & u_{\max} & \rightarrow & 0 & \rightarrow \\ u_v: & u_{\max} & \rightarrow & u_v^* & \rightarrow \\ t: & 0 & \rightarrow & t_{s_1} & \rightarrow . \end{array} \tag{3.33}$$

Remark 3.4. In the expressions (3.29)-(3.33) the instances when u_i can be singular are explicitly indicated. For example if u_i has the form specified in (3.32) then u_i is singular for all $t > \tau$. Conversely, if u_i has the form specified in (3.33) then u_i is never singular. In other words, form (3.33) not only specifies that $u_i(t) = 0$ for all $t > \tau$ but also that u_i is not singular.

Claim 3.3.7. *The optimal control is not of the form* (3.29).

Proof. If the optimal control is of the form (3.29) then it is also the optimal control for the isolation only model. Therefore, there exists a $t_s \in [0, T]$ such that

$$u_i(t) = \begin{cases} u_{\max} & t \in [0, t_s] \\ 0 & t \in (t_s, T], \end{cases}$$

is optimal. Now fix this choice of u_i. The next step is to show that there exists a u_v that gives a lower value for the cost function then $u_v \equiv 0$ does.

Suppose $t_s \neq T$ then, by the principle of optimality, it follows that

$$(u_i(t), u_v(t)) = \begin{cases} (u_{\max}, 0) & t \in [0, t_s) \\ (0, u_v^*) & t \in [t_s, T], \end{cases}$$

has a smaller cost then a control of the form (3.29).

Alternatively, suppose $t_s = T$, then the mixed model essentially becomes a vaccination only model with new parameter $\tilde{\mu} = \mu + u_{\max}$. Therefore, $u_v = u_v^*$ gives a lower value for the cost function than $u_v \equiv 0$ does. $\square$

Claim 3.3.8. *If the optimal control is of the form* (3.30) *then* $u_v^* \equiv u_{\max}$.

Proof. Suppose to the contrary that

$$u_v^*(t) = \begin{cases} u_{\max} & t \in [0, t_s] \\ 0 & t \in (t_s, T], \end{cases}$$

where $t_s < T$. Then it follows that

$$(u_i(t), u_v(t)) = \begin{cases} (0, u_{\max}) & t \in [0, t_s] \\ (u_i^*, 0) & t \in (t_s, T], \end{cases}$$

will give a lower value for the cost function then a control of the form (3.30) will. □

Claim 3.3.9. *The optimal control is not of the form* (3.30).

Proof. If the optimal control is of the form (3.30) then $\int_{t_0}^{T} u_i I \mathrm{d}t = 0 < w_{\max}$ and therefore $\lambda_w = 0$. Since u_i is not singular (see Remark 3.4), this implies that $\lambda_I(t) > 0$ for all $t > t_0$. Therefore, by Equation (3.27), $\dot{\lambda}_S(t) < 0$ for all $t > t_0$. This implies that $\lambda_S(t) > 0$ (since $\lambda_S(T) = 0$) for all $t < T$, but this gives a contradiction since if $\lambda_S(t) > 0$ then $u_v(t) = 0$. □

Claim 3.3.10. *If the optimal control is of the form* (3.32) *then* $\int_{t_0}^{t_{s_1}} u_{\max} S_{[u_{\max},u_{\max}]} \mathrm{d}t = z_{\max}$.

Proof. By Claim 3.3.5, if the optimal control has the form (3.32) then $u_s = u_i^*$. Therefore there exists a $t_{s_2} \in [t_{s_1}, T]$ such that

$$u_i(t) = \begin{cases} u_{\max} & t \in [t_0, t_{s_1}] \\ u_{\max} & t \in (t_{s_1}, t_{s_2}] \\ 0 & t \in (t_{s_2}, T] \end{cases}$$

is optimal. Assume that the number of people vaccinated at $t = t_{s_1}$ is not the maximal amount, that is assume $\int_{t_0}^{t_{s_1}} u_{\max} S_{[u_{\max},u_{\max}]} \mathrm{d}t < z_{\max}$. Suppose $t_{s_2} \neq T$ then it follows that

$$(u_i(t), u_v(t)) = \begin{cases} (u_{\max}, u_{\max}) & t \in [t_0, t_{s_1}] \\ (u_{\max}, 0) & t \in (t_{s_1}, t_{s_2}] \\ (0, u_v^*) & t \in (t_{s_2}, T], \end{cases}$$

has a smaller cost then a control of the form (3.32).

Alternatively, suppose $t_{s_2} = T$, then the mixed model essentially becomes a vaccination only model with new parameter $\tilde{\mu} = \mu + u_{\max}$. Therefore, $u_v = u_v^*$ is the optimal choice for u_v and so if $u_v = u_v^*$ has the form specified in (3.32) then it must be that $\int_{t_0}^{t_s} u_{\max} S_{[u_{\max}, u_{\max}]} \mathrm{d}t = z_{\max}$. $\square$

Claim 3.3.11. *If the optimal control is of the form* (3.33) *then* $\int_{t_0}^{t_s} u_{\max} I_{[u_{\max}, u_{\max}]} \mathrm{d}t = w_{\max}$.

Proof. The first step is to show that $\lambda_I(t) \leq 0$ for all $t \in [0, T)$. Suppose to the contrary that there exists a $t^* \in [0, T)$ such that $\lambda_I(t^*)$ is positive. Then, since $\dot{\lambda}_I \geq 0$, it must be that $\lambda_I(t) > 0$ for all $t \in J = [t^*, T]$. Therefore, since $\lambda_w \leq 0$, $u_i = 0$ on J. This implies, by Equation (3.27), that $\dot{\lambda}_S < 0$ on J and hence $\lambda_S > 0$ on J (since $\lambda_S(T) = 0$). Similarly, since $\lambda_S > 0$ on J, $u_v = 0$ on J and this implies, by Equation (3.27), that $\dot{\lambda}_I = 0$ on J. Letting $\dot{\lambda}_I = 0$ and rearranging Equation (3.28), gives

$$\lambda_I(\beta S - \mu) = (\lambda_0 + \lambda_S)\beta S. \tag{3.34}$$

Now both the right-hand side of Equation (3.34) and λ_I are positive on J. This implies that $S(t) > \frac{\mu}{\beta}$ for all $t \in J$. But this contradicts the fact that $S(T) < \frac{\mu}{\beta}$. This proves that $\lambda_I(t) \leq 0$ for all $t \in [0, T)$.

Now, $\dot{\lambda}_I \geq 0$, $\lambda_I \leq 0$ and u_i switches from $u_{\max}$ to 0 when $\lambda_I = \lambda_w$. Therefore if $\lambda_w = 0$ then the switch occurs while u_i is singular, since u_i is not singular (see Remark 3.4) it must be that $\lambda_w < 0$. $\square$

Lastly, notice that control form (3.31) covers the case when there are enough resources to isolate with maximal effort for the entire epidemic. This proves Theorem 3.3.

3.8 Discussion

The results presented in Section 3.4 contain many interesting features. Firstly, recall that Wickwire showed that the optimal isolation policy is either to isolate with maximal effort for the entire epidemic or to never isolate any infectives. In contrast, the optimal isolation policy presented in Theorem 3.1 is to isolate with maximal effort for the entire epidemic only if there are enough resources. If there are not enough resources, the optimal policy is any policy that uses all of the resources. In other words, the optimal isolation policy presented by Wickwire is characteristically different than the one presented here. Conversely, the optimal vaccination policies given in [41] and Theorem 3.2 have the same basic form. These differences are partially explained by the relations of PMP. A simple calculation shows that the basic form of the relations of PMP are the same for the case when the objective function includes a term linear in the control and the case when the total amount of resources are limited. Also, the optimal isolation policy is singular, so the relations of PMP are not sufficient to determine the optimal isolation policy. This means that the optimal isolation policies in [37] and Theorem 3.1 can be different. For the vaccination model however, the relations of PMP are necessary and sufficient.

Another interesting aspect of the isolation only model is that for certain parameter values, small changes in $w_{\max}$ can cause large changes in the value of the minimal cost. To see this, notice that setting $u_v \equiv 0$ and $u_i \equiv u_{\max}$ in (3.1) gives,

$$w_{[u_{\max}]}(T) = \int_{t_0}^{T} u_{\max} I_{[u_{\max}]} \mathrm{d}t = -\frac{u_{\max}}{\beta} \ln\left(\frac{S_{[u_{\max}]}(T)}{S_0}\right). \tag{3.35}$$

Rearranging Equation (3.35) gives, $S_{[u_{\max}]}(T) = S_0 \exp\left(\frac{-\beta w_{[u_{\max}]}(T)}{u_{\max}}\right)$ which expresses the final number of susceptibles for the case when $u_i \equiv u_{\max}$, in terms of $w_{[u_{\max}]}(T)$.

Figure 3.1 shows that if $S_{[u_{\max}]}(T) > \frac{\mu}{\beta}$, then as $w_{\max}$ switches from less than $w_{[u_{\max}]}(T)$ to greater than $w_{[u_{\max}]}(T)$, there will be a jump in $S_{[u_i^*]}(T)$. That is, if $S_{[u_{\max}]}(T) > \frac{\mu}{\beta}$, then there is a range of values of the final number of susceptible individuals, $S_{[u_i^*]}(T)$, that cannot be attained, regardless of the value of $w_{\max}$. Finally, it is worth pointing out that, whether or not $S_{[u_{\max}]}(T) > \frac{\mu}{\beta}$ can be determined directly from the model parameters (without resorting to simulations). To see this, notice that combining Equation (3.13) and Equation (3.35) shows that $S_{[u_{\max}]}(T) > \frac{\mu}{\beta}$ if and only if

$$-\frac{u_{\max}}{\beta}\ln\left(\frac{\mu}{\beta S_0}\right) > S_0 - \frac{\mu}{\beta} + \frac{\mu}{\beta}\ln\left(\frac{\mu}{\beta S_0}\right) - I_T + I_0. \tag{3.36}$$

The left panel of Figure 3.1 shows a scenario when Equation (3.36) is not satisfied and the right panel of Figure 3.1 shows a scenario when Equation (3.36) is satisfied.

The non-uniqueness of u_i^* when $w_{\max} < w_{[u_{\max}]}(T)$ means that epidemics with different dynamics can still optimize the cost function of Problem 3.1. The left panel in Figure 3.2 shows three different optimal isolation strategies for a case were $w_{\max} = 1450$ and $w_{[u_{\max}]}(T) = 2072$. In all three cases the total infectious burden is 9884.9 and the total number of infecteds is 4742.4.

The left panel in Figure 3.2 shows that very different dynamics result from these three different controls even though all three are optimal. For example the length of the epidemic described by the dotted curve is significantly shorter than the other two epidemics but it reaches a higher peak intensity (as measured by the maximum number of infecteds). Although the epidemics described by the dashed and solid curves are very similar, the dashed curve describing the number of infectives is distinctive because it has two peaks. Finally, notice that all three isolation strategies result in

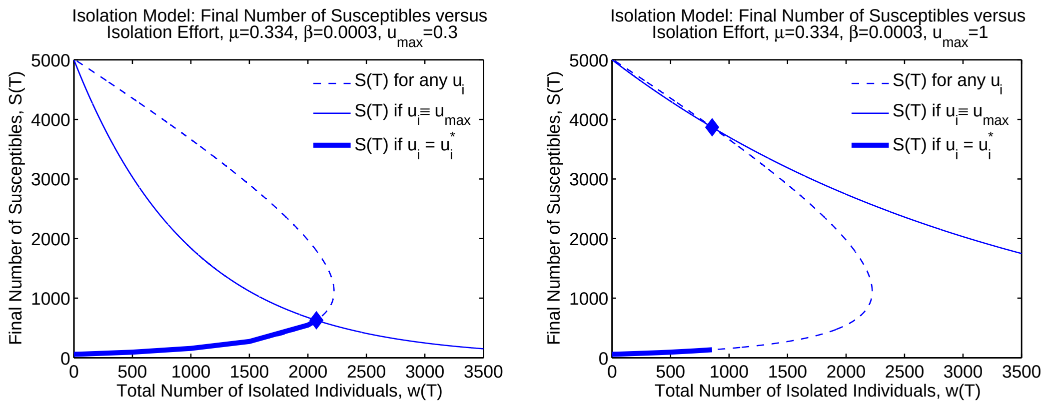

Figure 3.1: **Left Panel:** The final number of susceptibles depends only on the total number of isolated individuals (see Equation (3.13)). For parameter values, $I_0 = 10$, $S_0 = 5000$, $u_{\max} = 0.3$, $\mu = 0.334$ and $\beta = 0.0003$, the dashed curve describes this relationship. The diamond marks the point $(w_{[u_{\max}]}(T), S_{[u_{\max}]}(T))$. Since $u_i \equiv u_{\max}$ is the optimal solution for the unconstrained model and since minimizing the total number of infecteds is equivalent to maximizing $S(T)$, the diamond marks the largest attainable $S(T)$. The solid thick curve (which lies directly on top of the dashed curve) therefore denotes all possible optimal pairs $(w(T), S(T))$ for our specific choice of parameters. The point marked by the diamond is also the intersection of the dashed curve and the solid thin curve. The solid thin curve plots the final number of susceptibles under the assumption that $u_i \equiv u_{\max}$. $S(T) = \frac{\mu}{\beta}$ at the right-most point of the dashed curve. Since the point $(w(T), \frac{\mu}{\beta})$ on the solid thin curve lies to the left of the right-most point on the dashed curve, all optimal values of $S(T)$ must be less than $\frac{\mu}{\beta}$. **Right Panel:** This plot shows the same information as the plot on the left-hand side but for the case when $u_{\max} = 1$. For this case the intersection of the solid thin curve and the dashed curve occurs at an $S(T)$ larger than $\frac{\mu}{\beta}$. The solid thick curve shows all possible optimal pairs $(w(T), S(T))$ for cases when $w_{\max} < w_{[u_{\max}]}(T)$. For cases when $w_{\max} \geq w_{[u_{\max}]}(T)$, the optimal pair is $(w_{[u_{\max}]}(T), S_{[u_{\max}]}(T))$ (that is, the point marked by the diamond). Therefore, as $w_{\max}$ is increased past the value $w_{[u_{\max}]}(T)$ the final number of susceptibles jumps from a small value (less than 500) to a large value (greater than 3500).

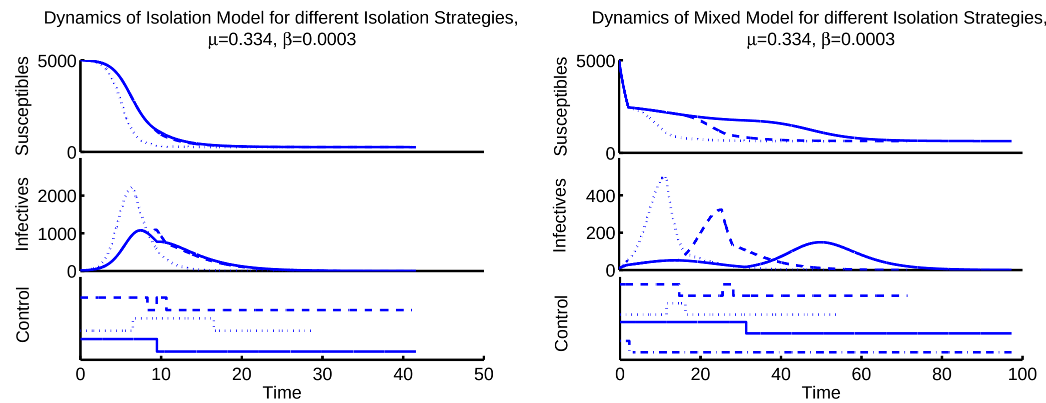

Figure 3.2: **Left Panel:** Dynamics of the isolation only model for three different, optimal isolation strategies. All three isolation strategies are purely bang-bang and switch between the values $u_{\min} = 0$ and $u_{\max} = 0.3$. The right plot shows the form of the different control strategies. The control strategies have been plotted offset from each other. The parameters used in these simulations are $I_0 = 10$, $S_0 = 5000$, $u_{\max} = 0.3$, $\mu = 0.334$, $w_{\max} = 1450$ and $\beta = 0.0003$. **Right Panel:** Dynamics of the mixed model for three different isolation strategies. All three isolation strategies are purely bang-bang and switch between the values $u_{\min} = 0$ and $u_{\max} = 0.3$. All three isolation strategies isolate 350 infecteds. The vaccination strategy is the same for all three cases. The right plot shows the form of the different control strategies. The control strategies have been plotted offset from each other. The solid curve and the dashed curve denote optimal strategies, the dotted curve shows a non-optimal strategy. The dash-dot curve in the right plot is the vaccination strategy, u_v. The parameters used in these simulations are $I_0 = 10$, $S_0 = 5000$, $u_{\max} = 0.3$, $\mu = 0.334$, $w_{\max} = 350$, $z_{\max} = 2500$ and $\beta = 0.0003$.

the same value $S(T) = 258$, as must be the case for all of them to be optimal.

Interestingly, this non-uniqueness means that policy makers have the freedom to optimize other cost functions in addition to (3.2). For example, as mentioned earlier, in this case it is possible to simultaneously minimize the total outbreak size, as well as the average number of infected individuals per unit time over the entire outbreak (which is a measure of its peakedness). The optimal strategy for doing so is to isolate individuals with maximal effort for as long as possible (Section 3.9).

The left panel of Figure 3.2 also emphasizes the difference between the optimal isolation only policy and the optimal vaccination only policy. Notice that even though the dotted isolation strategy does not begin until the epidemic is well under way, it still minimizes the total number of infectives (and therefore also the total infectious burden). This result is very different from the vaccination results.

The right panel in Figure 3.2 shows the results for three different isolation strategies for the mixed model. All three scenarios use an optimal vaccination strategy with $z_{\max} = 2500$. All of the isolation strategies isolate 350 infectives, but not all of these strategies are optimal. The dashed and the solid curves illustrate dynamics for optimal control strategies and the total number of infecteds for both of these control strategies is 1861. On the other hand, the control strategy denoted by the dotted line is not optimal (notice that the isolation control is off when the vaccination control is on) and consequently the total number of infecteds is larger, 1870.

It is also worth highlighting the significance of how disease characteristics place upper bounds on the control variables. As an example, suppose that a disease of interest is such that the maximum possible rate of isolation is $u_{\max} = 0.3$ as in Figure 3.2. In this case (assuming $w_{\max} = 1450$ as in Figure 3.2), the optimal isolation policy will result in a total outbreak size of 4742 individuals. This yields the counterintuitive finding that, although we have enough resources to isolate 1450 individuals, and although there are only 10 infected individuals that initiate the outbreak, we are nevertheless unable to use isolation to prevent the epidemic. More surprising still, our resources turn out to be insufficient for the outbreak that develops. In other words, even though only 10 individuals start the epidemic, and even though we have enough resources to isolate 1450 individuals, our lack of ability to quickly isolate these initial 10 infections results in an outbreak so large that our resources are then insufficient. This clearly highlights the importance of constraints on the rate of isolation (and vaccination) in addition to constraints on the total amount of resources that are available. Indeed, given the likelihood of a large amount of uncertainty in $u_{\max}$, these considerations provide a justification for using up all of the isolation resources as quickly as possible, despite the fact that ultimately this might not reduce total outbreak size any more that using all the resources in some other temporal fashion, just in case $u_{\max}$ is larger than we think.

The above observations suggest a number of interesting areas that would benefit from further analysis. For example, the non-uniqueness of the optimal isolation policy, coupled with the fact that various optimal strategies yield dramatically different dynamics, suggests that aspects of the outbreak "shape" should be included in the optimization problem, and not simply outbreak size or infectious burden. It is also interesting to note that in the General Problem discussed in this chapter, the values of $u_{\max}$, $z_{\max}$ and $w_{\max}$ are predetermined and fixed. It would be worthwhile to investigate the effects of varying these parameters and describing the interplay between $u_{\max}$ and the maximum possible number of isolated ($w_{\max}$) and vaccinated

($z_{\max}$) individuals. For example, if the optimal isolation policy is non-unique, then increasing the value of $u_{\max}$ (while keeping $w_{\max}$ fixed) will not change the value of the cost function provided that the new optimal control is still not unique. Conversely, increasing $u_{\max}$ (while keeping $z_{\max}$ fixed) will always improve the vaccination only policy. The effect of changing $u_{\max}$ for the mixed policy is not as easily understood nor is the effect of changing both $u_{\max}$ and $w_{\max}$ (or $z_{\max}$). Furthermore, it might also be interesting to examine what happens if the cost associated with using a control depends on the value of $w_{\max}$ (or $z_{\max}$) and not just on the total number of individuals that are isolated (or vaccinated). For instance, even if there is a fixed stockpile of vaccine, it is presumably more costly (in terms of needing nurses etc) to have a high vaccination rate than a low one at any given time. Lastly, the solution to the mixed model presented here assumes that the maximum allowed number of isolated and vaccinated individuals are constrained separately. It would be interesting to determine how to optimally divide a total resource pool into separate isolation and vaccination resource pools. This would be useful for example, when only the total amount of resources have been allocated and the policy-maker is free to choose how to divide these resources between vaccination and isolation.

3.9 Details of Proofs

The remaining details of the proofs of Theorems 3.1, 3.2 and 3.3 are presented here.

3.9.1 Details for Proof of Theorem 3.1

To derive Equation (3.13) from Equation (3.9) first note that Equation (3.8) can be rearranged to give (i) $-\dot{S} = \beta SI$ and (ii) $I = -\frac{\dot{S}}{\beta S}$. Substituting these two expressions into Equation (3.9) gives

$$\dot{I} = -\dot{S} + \frac{\mu}{\beta}\frac{\dot{S}}{S} - u_i I. \tag{3.37}$$

Rearranging Equation (3.37) and integrating from $t = t_0$ to $t = T$ results in

$$\int_{t_0}^{T} u_i I \mathrm{d}t = S_0 - S(T) + I_0 - I(T) + \frac{\mu}{\beta}\ln(\frac{S(T)}{S_0}). \tag{3.38}$$

The only detail that remains to be proved for Theorem 3.1 is that the optimal isolation strategy for the unconstrained problem is $u_i^* \equiv u_{\max}$.

Claim 3.1.1. *The optimal control for Problem 3.1 with* $w_{\max} = \infty$ *is* $u_i^* \equiv u_{\max}$.

Proof. The relations of PMP for Claim 3.1.1 are given by the relations of PMP in Section 3.5 with $\lambda_w = 0$. Substituting Equation (3.12) into Equation (3.11) gives,

$$\begin{aligned} \dot{\lambda}_S &= -(-\lambda_0 - \lambda_S + \lambda_I)\beta I \\ &= \dot{\lambda}_I \frac{I}{S} - \lambda_I(\mu + u_i)\frac{I}{S} \end{aligned} \tag{3.39}$$

The first step is to show that the optimal control is purely bang-bang (that is, it has no singular components). Since, by Equation (3.10), λ_I is a constant, if u_i is singular then it must be singular on the entire interval $[0, T]$. This implies, by Equation (3.39), that λ_S is constant and since $\lambda_S(T) = 0$ it must be that $\lambda_S \equiv 0$. Equation (3.11) then gives that $\lambda_0 = 0$. This is not possible since $(\lambda_0, \lambda_I(t), \lambda_S(t))$ must be nonzero for all $t \in [0, T]$. Therefore, u_i^* cannot be singular.

The optimal control will be determined once the sign of λ_I is determined. To determine the sign of λ_I, use the transversality condition $\lambda_S(T) = 0$. Since λ_I is a constant, Equation (3.12) gives that $\lambda_I = \frac{(\lambda_0+\lambda_S)\beta S}{\beta S-u-\mu} = \frac{\lambda_0 \beta S(T)}{\beta S(T)-u_i(T)-\mu}$. This implies that $\text{sign}(\lambda_I) = \text{sign}\left(S(T) - \frac{u_i(T)+\mu}{\beta}\right)$. By Equation 3.1, λ_I is negative if and only if $\dot{I}(T) < 0$. Since T is the smallest time that $I = 0.5$ (and $I(0) > 0.5$) it must be that $\dot{I}(T)$ is negative. Therefore, $u_i^* \equiv u_{\max}$. □

3.9.2 Details for Proof of Theorem 3.2

The only detail that remains to be proved for Theorem 3.2 is that the vaccination strategy $u_v \equiv 0$ is not optimal.

Claim 3.2.1. *The control $u_v \equiv 0$ is not a solution for Problem 3.2.*

Proof. From the relations of PMP given in Section 3.6, if $u_v^* \equiv 0$ then $\lambda_z \leq \lambda_S$. Also, since $u_v \equiv 0$ means that none of the vaccination resources are used, it must be that $\lambda_z = 0$. This implies that $\lambda_S \geq 0$.

Furthermore, by Equation (3.14), $u_v^* \equiv 0$ implies that $\dot{\lambda}_I \equiv 0$. Rearranging Equation (3.16) and setting $t = T$ gives $\lambda_I(\beta S(T) - \mu) = \lambda_0 \beta S(T)$. There are two possible scenarios:

(i) If $\lambda_0 = 0$ then $\lambda_I \equiv 0$ but since $\lambda_S(T) = 0$ this implies that the adjoint vector is zero at $t = T$. Therefore this scenario is not possible.

(ii) If $\lambda_0 = 1$ then $\lambda_I < 0$ (since $S(T) < \frac{\mu}{\beta}$) and this implies, by Equation (3.14), that $\dot{\lambda}_S > 0$. Therefore, since $\lambda_S(T) = 0$, it must be $\lambda_S(t) < 0$ for all $t < T$. But this contradicts the fact that $u_v^* \equiv 0$ implies that $\lambda_S \geq 0$.

□

3.9.3 Details for Proof of Theorem 3.3

We now prove the claims on pages 55-55 that were used to prove Theorem 3.3. Throughout this section the notation specified in Remark 3.3 will hold and the relations of PMP will be the ones given Section 3.7.

Claim 3.3.1. *If u_v is singular on some interval J then u_i is not singular on J and $u_i = 0$ on J.*

Proof. Suppose u_v is singular on some interval J, then $\dot{\lambda}_S = 0$ and therefore Equation (3.27) implies that $\lambda_I = \lambda_w \frac{u_i}{\mu+u_i}$. Now if $\lambda_w = 0$ then $\lambda_I = 0$. Since $\dot{\lambda}_S = 0$ it must be that $\lambda_I = \lambda_0 + \lambda_z$ and therefore $\lambda_0 = -\lambda_z$. Furthermore, since $T \in J$ (see Remark 3.5) it must be that $\lambda_z = 0$ (because $\lambda_S(T) = 0$) so the adjoint vector is identically zero on J. Since this is not possible it must be that $\lambda_w < 0$ and therefore $\lambda_w < \lambda_w \frac{u_i}{\mu+u_i} = \lambda_I$ so $u_i = 0$. □

Remark 3.5. If u_v is singular then $\dot{\lambda}_S = 0$. The only way for u_v to change from being singular to being nonsingular is if λ_I becomes nonconstant. But, by Equation (3.28), if both $\dot{\lambda}_S = 0$ and $\dot{\lambda}_I = 0$ then $\dot{\lambda}_I$ must remain zero. This implies that if u_v is singular on some maximal interval J then $T \in J$.

Claim 3.3.2. *If u_i is singular on an interval J, then $u_v = 0$ on J.*

Proof. If u_i is singular on some interval J, then $\dot{\lambda}_I = 0$ on J. This implies by Equation (3.27), that either u_v is singular or $u_v = 0$ on J. Since by Claim 3.3.1, u_i and u_v cannot be singular at the same time, it must be that $u_v = 0$ on J. □

Claim 3.3.3. *When u_i is singular, denote it by u_s. If u_i has a singular component* *then u_i has one of the following forms:*

$$\begin{array}{lcl} u_{\text{max}} & \longrightarrow & u_{\text{s}} \\ u_{\text{s}} & \longrightarrow & u_{\text{s}}. \end{array}$$

Proof. Pick any $\tau \in J$. Claim 3.3.3 follows directly from the following two observations. Firstly, notice that if u_i is singular on some interval J then, by Equation (3.27), $\dot{\lambda}_S \geq 0$ on J. This implies that $u_v(t) = 0$ for all $t > \tau$ and therefore, by Equation (3.27), $\dot{\lambda}_I(t) = 0$ for all $t > \tau$. That is if u_i is singular on J then $T \in J$. Secondly, by Equation (3.27), $\dot{\lambda}_I \geq 0$ (even when u_i is not singular). □

Claim 3.3.4. *If there exists a $t_s \geq 0$ such that u_i is constant for all $t \in (t_s, T]$ then $u_v(t) = u_v^*(t)$ for all $t \in (t_s, T]$.*

Proof. Once $t > t_s$ the mixed isolation-vaccination model becomes a vaccination only model and therefore the optimal u_v is the optimal control for the vaccination only model with parameters $\tilde{t}_0 = t_s$, $\tilde{z}_{\text{max}} = z_{\text{max}} - \int_{t_0}^{t_s} u_v S_{[u_i,u_v]} \mathrm{d}t$ and $\tilde{\mu} = \mu + u_i(T)$. □

Claim 3.3.5. *If there exists a $t_s \geq 0$ such that $u_v(t) = 0$ for all $t \in (t_s, T]$ then $u_i(t) = u_i^*(t)$ for all $t \in (t_s, T]$.*

Proof. Once $t > t_s$ the mixed isolation-vaccination model becomes an isolation only model with parameters $\tilde{t}_0 = t_s$, $\tilde{w}_{\text{max}} = w_{\text{max}} - \int_{t_0}^{t_s} u_i I_{[u_i,u_v]} \mathrm{d}t$ and $\tilde{\mu} = \mu$. Therefore the optimal u_i is the optimal control for the isolation only model. □

Claim 3.3.6. *Let J be any maximal interval such that*

(i) $u_i = u_{\text{max}}$ *on J and*

(ii) u_i is nonsingular on J.

If $T \notin J$ then $u_v = u_{\max}$ on J.

Proof. Conditions (i) and (ii) imply that $\lambda_I(t) < \lambda_w$ for all $t \in J$. Therefore, by Equation (3.27), $\dot{\lambda}_S(t) > 0$ for all $t \in J$. Therefore, while $t \in J$, u_v must have one of the following forms:

$$\begin{aligned} u_{\max} &\rightarrow u_{\max}, \\ u_{\max} &\rightarrow 0, \\ 0 &\rightarrow 0. \end{aligned}$$

Let $t_1 \in J$. Suppose $u_v(t) = 0$ for all $t \in \tilde{J} = \{t \mid t > t_1 \text{ and } t \in J\}$ then, by Equation (3.27), $\dot{\lambda}_I = 0$ on $\tilde{J}$. This implies that $\max\{\lambda_I(t) \mid t \in \tilde{J}\} < \lambda_w$, since if $\max\{\lambda_I(t) \mid t \in \tilde{J}\} \geq \lambda_w$ either (i) or (ii) would be violated. But if $\max\{\lambda_I(t) \mid t \in \tilde{J}\} < \lambda_w$ then either J is not maximal or $T \in J$. Since both of these possibilities lead to a contradiction, the only possible form for u_v is $u_v = u_{\max}$ on J. $\square$

Isolation Policy that Minimizes Average Number of Infectives

By introducing the variable $m = t$, the problem of minimizing the average number of infectives becomes the problem of finding the control u_i that maximizes $\frac{S(T)-S_0}{m}$ subject to the constraints that $I(T) \leq 0.5$ and $w(T) \leq w_{\max}$. The following relations are provided by PMP:

The Hamiltonian is

$$H(t) = \beta SI(\lambda_I - \lambda_S) + (\lambda_w - \lambda_I)u_i I - \lambda_I \mu I + \lambda_m = -\dot{\lambda}_I + \lambda_m = 0, \tag{3.40}$$

where the adjoint variables are defined by, $\dot{\lambda}_S = -\beta I(\lambda_I - \lambda_S)$, $\dot{\lambda}_m = 0$, $\dot{\lambda}_w = 0$ and

$\dot{\lambda}_I = -\beta I(\lambda_I - \lambda_S) - (\lambda_w - \lambda_I) + \lambda_I \mu$. The transversality conditions are

$$(\lambda_0, \lambda_S(T), \lambda_I(T), \lambda_w, \lambda_m) = \left(\lambda_0, \frac{\lambda_0}{m}, \lambda_I(T), \lambda_w, \lambda_0 \frac{S_0 - S(T)}{m^2}\right),$$

where $\lambda_I(T), \lambda_w \leq 0$. The optimal control therefore satisfies

$$u_i^* = \begin{cases} u_{\max}, & \lambda_w - \lambda_I > 0, \\ ?, & \lambda_w - \lambda_I = 0, \\ 0, & \lambda_w - \lambda_I < 0. \end{cases}$$

The first step is to show that the optimal control is strictly bang-bang. Suppose u_i^* is singular on an interval J, then $\lambda_w - \lambda_I = 0$ on J. This implies that $\dot{\lambda}_I = 0$ on J. By Equation (3.40), this implies that $\lambda_m = 0$ which (by the transversality conditions) implies that $\lambda_0 = 0$. Furthermore, since $H = 0$ and $\dot{\lambda}_m = 0$ this implies that $T \in J$. Substituting $(\lambda_0, \lambda_S, \lambda_I, \lambda_w, \lambda_m)|_{t=T} = (0, 0, \lambda_w, \lambda_w, 0)$ into $H = 0$ gives that $\lambda_w = 0$ and therefore the optimal control must be strictly bang-bang (since the adjoint vector must always be nonzero).

Since $\lambda_m \geq 0$, Equation (3.40), implies that $\dot{\lambda}_I \geq 0$ and therefore the switch function is non-increasing. This implies that the optimal control is either $u_i^* \equiv u_{\max}$, $u_i^* \equiv 0$ or u_i^* switches once from $u_{\max}$ to 0. To further constrain the form of the optimal control, consider the case when $\lambda_w = 0$. Since $\lambda_I(T) \leq 0$ and $\dot{\lambda}_I > 0$ it must be that $\lambda_I(t) < 0$ for all $t < T$ and therefore $\mathrm{SF}(t) \geq 0$. Therefore, if $\lambda_z = 0$ the optimal control is to isolate with maximal effort for the entire epidemic.

Now consider the case when $\lambda_w < 0$. If $\lambda_w < 0$ then all of the control resources are used and therefore the optimal control is to isolate with maximal effort until all of the resources are used.

Combining both cases, gives that the optimal control is to isolate with maximal effort until either all of the resources are used or the epidemic is over.

3.10 Simulations for More Detailed Models

Here we consider two more detailed versions of model (3.1). The first version includes more detailed isolation dynamics and allows for the possibility that isolated individuals may still transmit infection. The more detailed isolation model is described by the following set of equations:

$$\begin{aligned}
\dot{S} &= -(\beta I + \beta_J J)S - u_v S, \\
\dot{I} &= (\beta I + \beta_J J)S - (\mu + u_i)I, \\
\dot{J} &= u_i I - \mu J,
\end{aligned} \tag{3.41}$$

where S, I and J are the numbers of susceptible, infected and isolated hosts, β is the transmission constant for infected hosts, β_J is the transmission constant for isolated hosts, μ is the per capita loss rate of infected individuals through both mortality and recovery, and u_i is the per capita rate of isolation.

Perhaps the most important result for the isolation model is that if there are not enough resources to isolate for the entire outbreak then any isolation strategy that uses the maximum amount of resources is optimal. Figure 3.3 suggests that this result also holds for the more detailed isolation model. More specifically, if E_i^* is the number of individuals isolated when $u_i \equiv u_{\max}$, then Figure 3.3 suggests that all isolation strategies that isolate $E_i < E_i^*$ result in the same outbreak size. It is important to emphasize that these results are only suggestive and that much more detailed analysis is required in order to make a stronger statement.

The second version includes more detailed vaccination dynamics and allows for the possibility that vaccinated individuals may still contract infection. The more detailed vaccination model is described by the following set of equations:

$$\begin{aligned}
\dot{S} &= -(\beta I + \beta_v I_v)S - u_v S, \\
\dot{S}_v &= u_v S - (1-r)(\beta I + \beta_v I_v)S_v, \\
\dot{I} &= (\beta I + \beta_v I_v)S - (\mu + u_i)I, \\
\dot{I}_v &= (1-r)(\beta I + \beta_v I_v)S_v - (\mu + u_i)I_v, \qquad (3.42)
\end{aligned}$$

where S and I are the numbers of susceptible and infected hosts, S_v and I_v are the numbers of vaccinated susceptible and vaccinated infected hosts, r represents the reduction in susceptibility of vaccinated individuals, β is the transmission constant for infected hosts, β_v is the transmission constant for vaccinated infected hosts, μ is the per capita loss rate of infected individuals through both mortality and recovery, and u_v is the per capita rate of vaccination. Perhaps the most important result for the vaccination model is that it is optimal to begin vaccination immediately and to vaccinate for the entire epidemic. Figure 3.4 suggests that this result also holds for the more detailed vaccination model. More specifically, if E_v^* is the number of individuals vaccinated when $u_v \equiv u_{\max}$, then Figure 3.4 shows that vaccination strategies that vaccinate the same number of individuals result in different outbreak sizes. Furthermore, of the strategies considered, it is always best to start vaccinating immediately and to continue vaccinating for longer. It is important to emphasize that these results are only suggestive and that much more detailed analysis is required in order to make

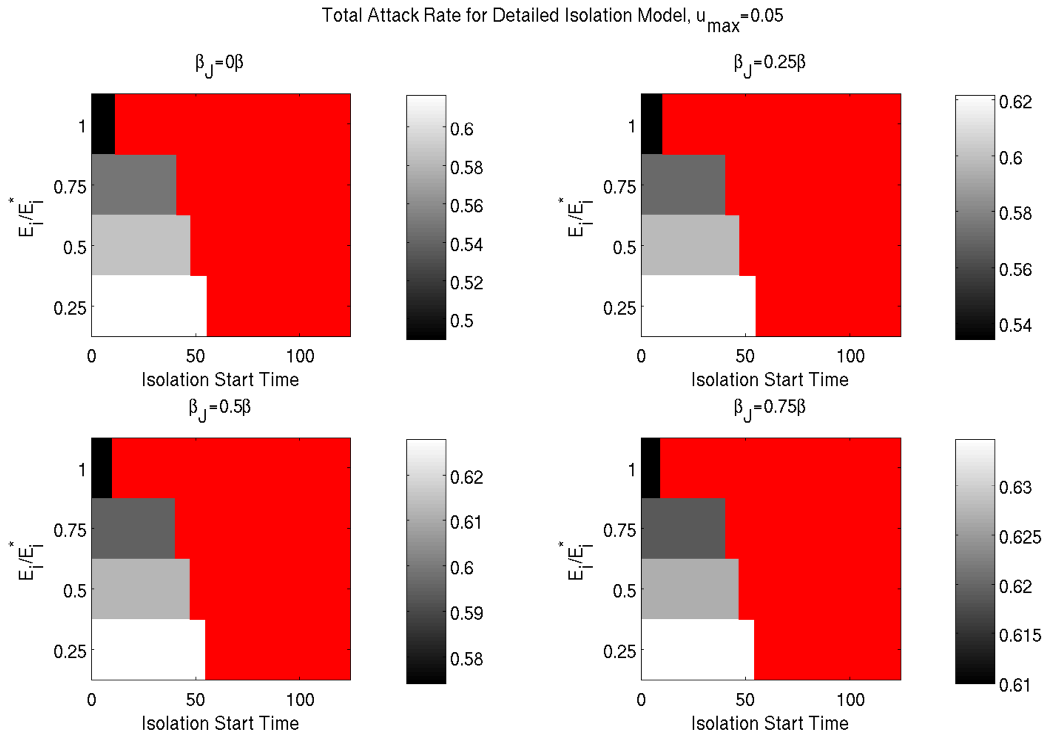

Figure 3.3: Results suggesting that if there are not enough resources to isolate for the entire outbreak then any isolation strategy that uses the maximum amount of resources is optimal. Each panel corresponds to a different value of β_J. The upper left panel corresponds to $\beta_J = 0$ and therefore corresponds to the situation when isolation is completely effective. The bottom right panel corresponds to $\beta_J = 0.75\beta$ and therefore corresponds to the situation when isolation is not very effective. E_i^* denotes the total number of isolated individuals when isolation starts immediately and continues for the entire outbreak with maximal effort. Each panel shows the attack rate as a function of total isolated (normalized by E_i^*) and isolation start time. Each point on the panels was generated by beginning isolation at the time indicated on the horizontal axis and continuing to isolate with maximal effort until the number isolated is the value indicated on the vertical axis. Once the total isolated has reached the number indicated on the vertical axis, no more hosts are isolated. If the outbreak ends before enough individuals have been isolated, the corresponding point on the plot is colored red. Thus red areas on the panels should be ignored. All four panels suggest that all isolation strategies that isolate $E_i < E_i^*$ result in the same outbreak size (ie., the shade of gray is constant for all isolation start times that correspond to the same value of E_i). Parameters used are $S_0 = 100000$, $I_0 = 10$, $\mu = \frac{1}{3.3}$, $\beta = \frac{1.6\mu}{S_c a0}$, $\beta_v = 0.8\beta$, $u_{\max} = 0.05$.

a stronger statement.

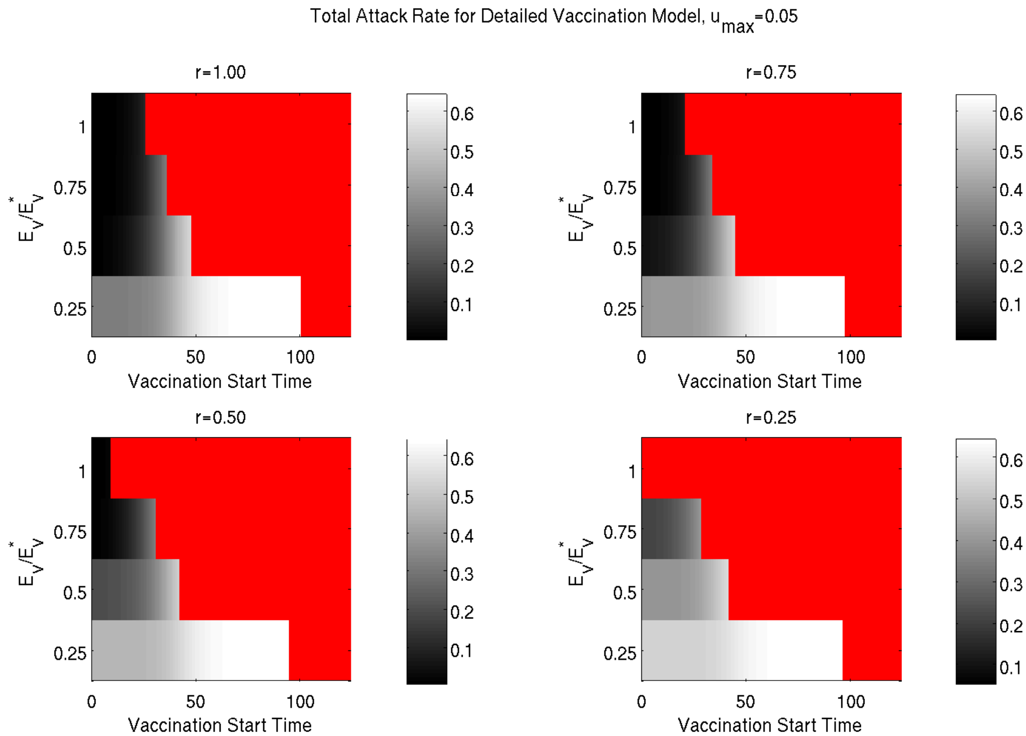

Figure 3.4: Results suggesting that it is optimal to start vaccinating immediately and to vaccinate for the entire outbreak.Each panel corresponds to a different value of r. The upper left panel corresponds to $r = 1$ and therefore corresponds to the situation when vaccination is completely effective. The bottom right panel corresponds to $r = 0.25$ and therefore corresponds to the situation when vaccination is not very effective. E_v^* denotes the total number of vaccinated individuals when vaccination starts immediately and continues for the entire outbreak with maximal effort. Each panel shows the attack rate as a function of total vaccinated (normalized by E_v^*) and vaccination start time. Each point on the panel was generated by beginning vaccination at the time indicated on the horizontal axis and continuing to vaccinate with maximal effort until the number vaccinated is the value indicated on the vertical axis. Once the total vaccinated has reached the number indicated on the vertical axis, vaccination is stopped. If the outbreak ends before enough individuals have been vaccinated, the point is colored red. Thus red areas on the panels should be ignored. All four panels suggest that starting vaccination immediately is best (the smaller the vaccination start time the smaller the attack rate, or in other words the shade of gray lightens as we move from left to right). All four panels show that, of the vaccination strategies represented on the panels, vaccinating for the entire outbreak results in the lowest attack rate. For the bottom right panel, of the vaccination strategies considered the only one that vaccinates E_v^* individuals is the one that the vaccinates for the entire outbreak. Therefore, the entire upper row (except the point corresponding to a vaccination start time of $t = 0$) is red. Parameters used are $S_0 = 100000$, $I_0 = 10$, $\mu = \frac{1}{3.3}$, $\beta = \frac{1.6\mu}{S_0}$, $\beta_v = 0.8\beta$, $u_{\max} = 0.05$.

Chapter 4

Optimal Antiviral Treatment Strategies and the Effects of Resistance

4.1 Abstract

Recent pandemic planning has highlighted the importance of understanding the effect that widespread antiviral use will have on the emergence and spread of resistance. A number of recent studies have determined that, if resistance to antiviral medication can evolve, then deploying treatment at a less than maximum rate often minimizes the outbreak size. This finding, however, involves the assumption that treatment levels remain constant during the entire outbreak.

Using optimal control theory, we address the question of optimal antiviral use by considering a large class of time-varying treatment strategies. We prove that, contrary to previous results, it is always optimal to treat at the maximum rate provided that

this treatment occurs at the right time. In general the optimal strategy is to wait some fixed amount of time and then to deploy treatment at the maximum rate for the remainder of the outbreak. We derive analytical conditions that characterize this optimal amount of delay. Our results show that it is optimal to start treatment immediately when one of the following conditions holds: (i) immediate treatment can prevent an outbreak; (ii) the initial pool of susceptibles is small; or (iii) when the maximum possible rate of treatment is low, such that there is little de novo emergence of resistant strains. Finally, we use numerical simulations to verify that the results also hold under more general conditions.

4.2 Introduction

The current interest in influenza pandemics has emphasized the importance of understanding the implications of drug-resistance for different types of public health interventions. Recent studies have considered the consequences of drug resistance on the effectiveness of several intervention strategies, including drug prophylaxis and treatment, vaccination, nondrug interventions, as well as combinations of interventions and multi-drug therapy [5, 12, 14, 16–18, 20, 22, 28–30, 32, 33, 36, 38, 39, 43]. Many of these studies have revealed interesting and unexpected behaviour. In particular, numerous studies have now demonstrated that, with respect to drug treatments, there is sometimes an intermediate optimal level of treatment that minimizes the total outbreak size [22, 28, 32, 33, 36].

This result is somewhat counterintuitive, and arises from a trade-off between the costs and benefits of treating infected individuals. In fact, treatment has three effects on the total outbreak size: (i) treatment is beneficial because it suppresses the spread of the sensitive strain; (ii) treatment is costly because it leads to the de novo appearance of resistant infections; and, (iii) treatment is costly because suppression of the sensitive strain frees up susceptible hosts that can then be infected by the resistant strain [28]. The treatment level that minimizes the total outbreak size should thus strike the optimal balance among these three effects.

These findings clearly reveal that the prospect of pathogen evolution during a disease outbreak can have a significant impact on the design of optimal intervention strategies. Nevertheless, it is important to note that most of these results make a key implicit assumption; namely that, whatever treatment level is chosen, it must remain constant during the entire course of the outbreak. Interesting exceptions are the studies by Moghadas et al. [32, 33], where numerical examples are used to illustrate that an even smaller total outbreak size can sometimes be obtained by switching from one level of treatment to another once the outbreak is underway.

Given the above collection of results, it is clear that the optimal treatment strategy during an outbreak is not yet completely understood. In this chapter we address this issue through the use of optimal control theory [13, 26, 27, 35]. In particular, we consider a model in which we allow for arbitrary, time-varying, treatment strategies during an outbreak, and we derive analytical expressions that characterize the form of the optimal strategy. We prove that, contrary to the implications of previous studies, it is always optimal to use maximum treatment levels *provided that this treatment occurs at the right time*. In general the optimal strategy is to wait some fixed period of time once the outbreak has begun, and then to deploy treatment at a maximum level. We show that the optimal delay is one that balances the above-mentioned costs and benefits. Finally, we explore extensions of the model to more complex scenarios in order to illustrate the robustness of our conclusions.

The model, depicted in Figure 4.1 (Panel A), describes the basic transmission dynamics of an infectious disease and includes the effects of treatment, as well as the presence of a strain resistant to treatment. An individual can be susceptible (S), infected with a sensitive strain (I), infected with a sensitive strain and treated (T), or infected with a resistant strain (R). Susceptible individuals become infected with the sensitive strain through contact with individuals in the I and T compartments (with contact parameters β_I and β_T respectively), and become infected with the resistant strain through contact with individuals in the R compartment (with contact parameter β_R). Individuals infected with the sensitive strain can move into the treated class at some (time-varying) per capita rate, $u(t)$, where u is the treatment strategy of interest. Treated individuals can develop resistance at some constant per capita rate, ν, and finally, individuals in each of the three infected classes can leave the system, through either death or recovery with immunity, at constant per capita rates μ_I, μ_T, and μ_R (also see Supplemental Information, Section 4.6.1).

4.3 Mathematical Model

The objective of the analysis is to determine the optimal time-varying treatment strategy, u, that minimizes the total attack ratio (i.e., minimizes the attack ratio of the sensitive strain plus the attack ratio of the resistant strain), given by

$$A = \frac{S(t_0) - S(t_f)}{S(t_0)},$$

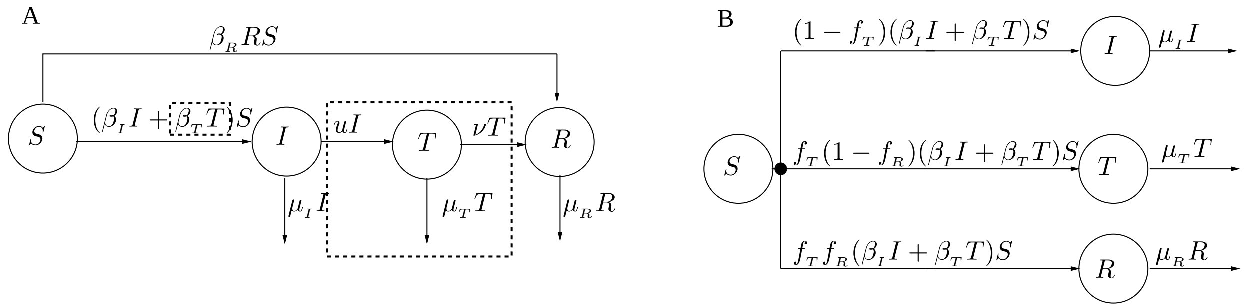

Figure 4.1: **Model structure. Panel A:** Schematic for the treatment model that is used to motivate model (4.1). An individual can be susceptible (S), infected with a sensitive strain (I), infected with a sensitive strain and treated (T), or infected with a resistant strain (R). The dynamics of the treated class have been enclosed in dashed boxes to emphasize the difference between this figure and the equations for model (4.1). All analytic results are derived using model (4.1) which does not include the dynamics of the treated class. **Panel B:** Schematic for the detailed model. Using numerical simulations the analytic results for model (4.1) have been extended to this more detailed model that includes the dynamics of the treated class.

which is equivalent to minimizing the total outbreak size (and to maximizing the final number of susceptibles). The quantities t_0 and t_f are the start and end times of the outbreak respectively. We say that the outbreak has ended as soon as the total number of all disease transmitting individuals is equal to 1. In other words, for the model depicted in Figure 4.1 (Panel A), $t_f = \min\{t \mid I(t) + T(t) + R(t) \leq 1\}$. By defining t_f in this way, we avoid the situation where an unrealistically small number of infecteds (namely less than one infected) initiates an outbreak. The rationale for this definition of t_f is further detailed in the Supplemental Information, Section 4.6.2 and in [23].

The treatment level, u, can vary through time, and can take on any value from 0 to some maximum value, $u_{\max}$. The upper bound, $u_{\max}$, reflects the fact that, regardless of how much we strive to increase the rate of treatment, it will typically be unavoidable that infected individuals spend some amount of time circulating in the population before they are treated. This will occur due to constraints on treatment delivery (e.g., having a finite number of health professionals available to deliver treatment), as well as constraints imposed by the biology of the disease (e.g., if there is an asymptomatic stage).

The above model is qualitatively similar to those examined by [22, 28, 33, 36] but we make a further pair of simplifying assumptions to facilitate mathematical analysis: (i) that treatment is perfectly effective, meaning that essentially no transmission occurs from treated individuals; and, (ii) that the time dynamics of the treated class are fast relative to those of the other classes. In this case we can use the quasi-equilibrium value of T to simplify the model (see Supplemental Information, Section 4.6.1). This simplification removes the highlighted elements of the flow diagram in Figure 4.1 (Panel A), and yields the following equations:

$$
\begin{aligned}
\dot{S} &= -\beta_I IS - \beta_R RS, \\
\dot{I} &= \beta_I IS - (\mu_I + u)I, \\
\dot{R} &= \beta_R RS - \mu_R R + \kappa u I,
\end{aligned}
\tag{4.1}
$$

where κ denotes the probability that a treated individual develops resistance.

Although the above simplifying assumptions might not be particularly realistic for many situations, our primary aim is to develop a broad conceptual understanding of the optimal treatment strategy during an outbreak. Moreover, these assumptions do not appear to qualitatively alter the results, as evidenced by the numerical simulations presented later. In fact, these simplifications primarily accentuate the three key effects of treatment that were discussed in the introduction. First, model (4.1) assumes that treatment completely suppresses the spread of the sensitive strain and therefore it maximizes the first effect of treatment (i.e., the suppression of sensitive infections). Second, model (4.1) assumes that, if de novo resistance appears at all, it arises immediately upon treatment because the dynamics of the treated class are assumed to be fast relative to the other classes. Third, model (4.1) also removes the impact of treated individuals on the depletion of susceptible hosts, therefore maximizing the third effect of treatment (i.e., the freeing up of susceptible hosts for infection by resistant strains).

4.4 Results

To facilitate the understanding of our results we first define some important quantities and briefly describe some of the basic features of the optimal treatment strategy. We then give a complete characterization of the optimal treatment strategy in two steps. First, we consider a special case (Situation 1 below) in which we can use standard techniques from calculus to gain an intuition for the results. Second, we characterize the optimal treatment strategy in general (Situation 2 below). Analysis of the latter case requires techniques from optimal control theory, and these details are presented in the Supplemental Information (Section 4.6).

The optimal treatment strategy for model (4.1) is characterized by the following three quantities:

(i) $R_I = \frac{\beta_I}{\mu_I} S(t_0) = \frac{S(t_0)}{S_I}$,

(ii) $R_{R} = \frac{\beta_R}{\mu_R} S(t_0) = \frac{S(t_0)}{S_R}$, and

(iii) $R_{IR} = \left(\frac{\beta_I}{\mu_I + u_{\max}} + \frac{\beta_R}{\mu_R} \frac{\kappa u_{\max}}{\mu_I + u_{\max}} \right) S(t_0)$.

Here, R_I and R_R are the basic reproduction numbers of the sensitive and resistant strain, respectively [7, 15, 25]. The quantities $S_I = \frac{\mu_I}{\beta_I}$ and $S_R = \frac{\mu_R}{\beta_R}$ represent threshold parameters, and denote the number of susceptibles above which a sensitive and resistant outbreak can occur, respectively. To understand the constant R_{IR} consider the possible paths that a new sensitive infection can take. If the population is receiving treatment (i.e., $u \neq 0$), then an initially sensitive infection may develop resistance and so some initially sensitive infections will generate both sensitive and resistant infections during their lifetime. The constant R_{IR} can be interpreted as the total expected number of secondary infections (both sensitive and resistant) caused by an individual initially infected with the sensitive strain, in a wholly susceptible population that is receiving maximum treatment (i.e., $u \equiv u_{\max}$). Specifically, the first term represents the number of infections generated while infected with the sensitive strain, and the second term represents the probability of developing resistance, multiplied by the number of infections generated while infected with this newly evolved resistant strain.

A few more general observations will be helpful before stating the results. First, R_{IR} can be expressed as $R_{IR} = R_I(1 - \chi) + \chi \kappa R_R$, where $\chi = \frac{u_{\max}}{(\mu_I + u_{\max})}$ is the probability of an infected individual receiving treatment when the treatment effort is $u \equiv u_{\max}$. This expresses R_{IR} as a mixture of the two reproduction numbers, R_I and R_R. Second, because the number of susceptibles always decreases through time, it is sometimes easier to express the optimal treatment strategy as a function of the size of the susceptible pool rather than as a function of time per se. This is also more useful from a health policy point of view because "number of susceptibles" is a much more meaningful (and more easily measured) quantity than "time since start of outbreak". Thus, although in the sequel we will discuss "treatment start time" and "delay", these will sometimes be measured in terms of the corresponding number of susceptible individuals. For reference all symbols have been defined in Table 4.1.

4.4.1 The Optimal Treatment Strategy

The optimal treatment strategy is to delay treatment until some fixed time τ^* (possibly $\tau^* = t_0$) and then to treat maximally for the remainder of the epidemic. Thus, higher treatment levels are always better provided that the onset of treatment is delayed by the optimal amount.

4.4.2 The Optimal Amount of Delay

To simplify the conditions for the optimal delay, suppose that $R_I \geq R_R$ (the resistant strain suffers a fitness cost in the absence of treatment) and $\kappa < 1$ (not all treated infections develop resistance; the complete analysis is presented in the Supplemental Information, Section 4.6.4). In this case, the optimal delay can be characterized in terms of a critical number of susceptible hosts, defined by

$$S_{\min} = S_R + \frac{\kappa}{1-\kappa}(S_R - S_I).$$

The optimal delay is achieved by balancing the effect that delay has on the sensitive and resistant outbreaks, and this is best understood by considering two separate situations: (i) the maximum treatment level is very large ($u_{\max} \to \infty$), and (ii) the maximum treatment level is not very large ($u_{\max}$ is finite).

Situation 1: If the maximum treatment level is very large, then some insight into the optimal treatment strategy can be gained using basic calculus. In particular, if the maximum level of treatment is used then the onset of treatment effectively ends the sensitive epidemic and initiates the start of the resistant epidemic. This is because at the onset of treatment all sensitive-infected individuals either enter the removed compartment (i.e., they are treated and no longer transmitting the infection) or the resistant compartment (i.e., treatment causes them to immediately develop resistance). Therefore, before treatment begins model (4.1) simplifies to,

$$\begin{aligned} \dot{S} &= -\beta_I I S, \\ \dot{I} &= \beta_I I S - \mu_I I, \end{aligned} \tag{4.2}$$

with initial conditions $(S(t_0), I(t_0))$. Once treatment begins at time $t = \tau$ all of the sensitive-infected individuals either (i) immediately become resistant or (ii) immediately stop transmitting the infection. Thus for $t > \tau$ model (4.1) simplifies to,

$$\begin{aligned} \dot{S} &= -\beta_R R S, \\ \dot{R} &= \beta_R R S - \mu_R R, \end{aligned} \tag{4.3}$$

with $(S(\tau), R(\tau)) = (S(\tau), \kappa I(\tau))$. Hence, in this special case, determining the optimal treatment strategy amounts to determining the optimal time to switch from model (4.2) to model (4.3).

Integrating and rearranging the equations for model (4.2) and model (4.3) gives (see Supplementa1 Information, Section 4.6.3 for details):

$$S_R \ln S(t_f) - S(t_f) = (1-\kappa)\left(S_{\min} \ln S(\tau) - S(\tau)\right) + C \tag{4.4}$$

where $C = \kappa(S_I \ln S_0 - S_0 - I_0) + R(t_f)$. Now since, by definition, the outbreak ends as soon as the total number of infecteds is less than or equal to 1 there are two possibilities with respect to the optimal delay:

(A) *Treating immediately can prevent the outbreak:* If $\kappa I(t_0) < 1$ then starting treatment at t_0 prevents the outbreak (i.e., $R(\tau) = \kappa I(\tau) < 1$). Clearly, this is the best possible scenario, and therefore $\tau^* = t_0$. Notice that this situation is closely related to the idea of "containment" discussed in [28].

(B) *Treating immediately cannot prevent the outbreak:* If the outbreak cannot be prevented through treatment, then any treatment strategy will result in $S(t_f) \leq S_R$ (i.e., the outbreak will not end until the pool of susceptibles is diminished below that necessary to sustain an resistant outbreak). Therefore the treatment start time that maximizes the left-hand side of Equation (4.4) will also maximize $S(t_f)$ (and hence minimize the total attack ratio). In other words it is optimal to delay treatment until the number of susceptible hosts has declined to $S(\tau^*) = S_{\min}$. If the initial number of susceptible hosts is smaller than this threshold, then treatment should begin immediately.

Situation 2: If the maximum treatment level is bounded, then there will be some temporal overlap in the sensitive and resistant epidemics once treatment has been initiated. In this situation, Pontryagin's Maximum Principle can be used to determine the optimal (time-varying) treatment strategy (see Supplemental Information, Section

4.6.4 for details). There are, again, two possibilities with respect to the optimal delay:

(A′) *Treating immediately can prevent the outbreak:* If starting treatment at t_0 prevents the outbreak from occurring then this is clearly the best possible scenario and so $\tau^* = t_0$. As with Situation 1, this situation is closely related to the idea of "containment" discussed in [28] and is contingent on how t_f is defined (i.e., for model (4.1) we have that $t_f = \min\{t \mid I(t) + R(t) \leq 1\}$).

(B′) *Treating immediately cannot prevent the outbreak:* If scenario (A′) does not hold then it may be optimal to delay the onset of treatment. In particular the threshold number of susceptibles $S_{\min}$ now provides an upper bound on the optimal delay. Thus, if the initial number of susceptible hosts is smaller than this threshold, then treatment should begin immediately. On the other hand, if $S(t_0)$ is initially larger than $S_{\min}$, then the optimal delay is no larger than the time it takes for the number of susceptible hosts to decline to $S_{\min}$. There is one other possibility, however, that overrides these conditions: if $R_{IR} > R_R$ then it is optimal to start treatment immediately.

Note that there are two major differences between the results for Situation 1 and Situation 2. First, when maximum treatment levels are bounded (Situation 2) the critical threshold, $S_{\min}$, provides only an upper bound for the optimal delay. Second, in Situation 2 there is an additional condition under which immediate treatment is optimal; namely, when $R_{IR} > R_R$. Intuitively, if de novo resistance has a negligible effect on the outbreak then it seems reasonable that treatment should proceed as though there is no de novo resistance. The condition that treatment should begin immediately if $R_{IR} > R_R$ actually highlights this intuition and emphasizes that the quantity that measures the effect of de novo resistance is the rate of de novo resistance ($\kappa u I$) and not the probability of de novo resistance (κ). This can be seen by rearranging $R_{IR} > R_R$ and multiplying both sides by κI to produce a relationship involving the rate of de novo resistance:

$$\kappa u_{\max} I < \frac{\kappa}{1-\kappa}\left(\frac{\mu_R}{\beta_R} - \frac{\mu_I}{\beta_I}\right)\beta_I I. \tag{4.5}$$

Hence, if the rate of de novo resistance is small then treatment should proceed as though there is only a treatment-sensitive strain (i.e., if equation (4.5) holds then treatment should start immediately).

4.4.3 Interpretation

We have shown that, for a simple model that includes the evolution of resistance, the optimal treatment strategy is to delay treatment for a specific amount of time (possibly no time) and then to treat with maximum effort for the remainder of the epidemic. If an outbreak is unavoidable, the optimal amount of delay is determined by balancing the benefits and costs of delay, and this trade-off can be best understood by again considering the case when the maximum treatment level is very large.

When the maximum treatment level is very large (i.e., Situation 1), the marginal change in the attack ratio that comes from increasing the amount of delay before treatment is (see Supplemental Information, Section 4.6.3 for details)

$$
\begin{aligned}
\frac{\mathrm{d}A}{\mathrm{d}\tau} &= \frac{\mathrm{d}A_I}{\mathrm{d}\tau} + \frac{\mathrm{d}A_R}{\mathrm{d}\tau}, && (4.6)\\
&= \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{S(t_0) - S(\tau)}{S(t_0)}\right) + \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{S(\tau) - S(t_f)}{S(t_0)}\right), && (4.7)\\
&= \alpha[S_R - S(\tau)] - \alpha\kappa[S_I - S(\tau)], && (4.8)
\end{aligned}
$$

where A, A_I, A_R denote the total, sensitive and resistant attack ratios respectively, τ is the treatment start time, and $\alpha = \frac{S(t_f)\beta_I I(\tau)}{S(t_0)(S_R - S(t_f))} > 0$.

Equation (4.6) emphasizes that the total attack ratio can be decomposed into a sensitive attack ratio and a resistant attack ratio. Furthermore, since for Situation 1 the sensitive and resistant outbreaks are temporally separated (because the resistant strain emerges only after treatment begins) the sensitive attack ratio includes all new infections that occur before the treatment start time τ and the resistant attack ratio includes all new infections that occur after the treatment start time τ (see Equation (4.7)). The two terms in Equation (4.8) represent the benefit and cost of delay, respectively. The first term represents the benefit of a delay that comes from a reduction in the size of the susceptible pool at the start of the resistant-strain epidemic. From the perspective of this benefit term only, it is best to delay treatment until $S(\tau) = S_R$, at which point there are no longer enough susceptible individuals to sustain a resistant outbreak. The second term represents the cost of delay that comes from an increase in the number of de novo resistant infections once treatment begins. From the perspective of this cost term only, it is best to treat immediately because waiting longer results in a larger pool of infected individuals which then have the potential to develop de novo resistance once treatment begins. The optimal amount of

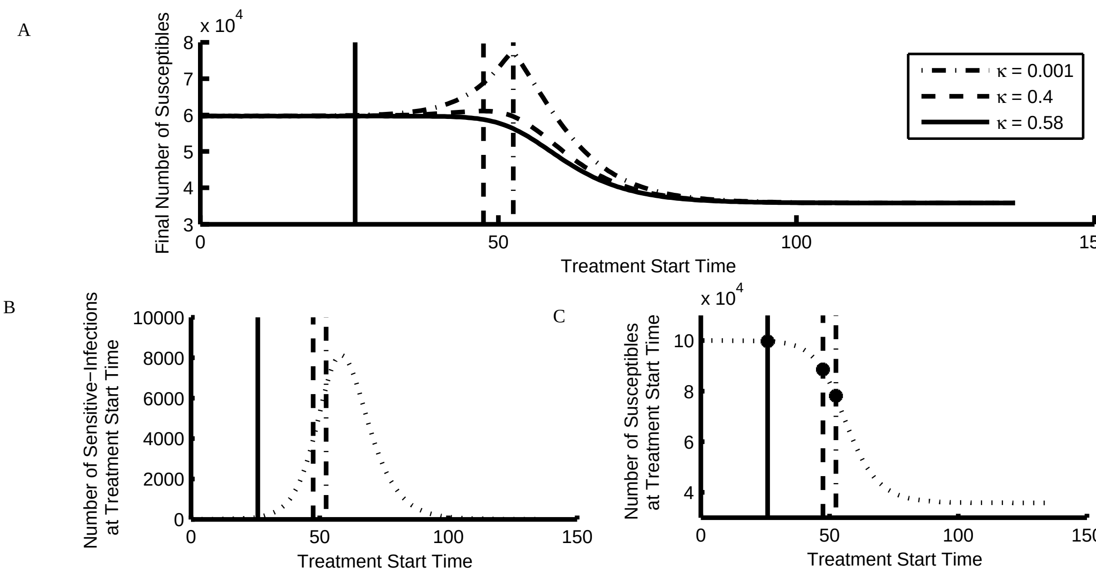

Figure 4.2: **Effect of treatment start time on total attack rate (Situation 1.B).** The vertical lines indicate the optimal treatment start time τ^* for a large κ (solid line), an intermediate κ (dashed line) and a small κ (dot-dashed line). **Panel A**: The final number of susceptibles as a function of treatment start time for a large κ (solid curve), an intermediate κ (dashed curve) and a small κ (dot-dashed curve). **Panel B**: The number of sensitive-infections as a function of treatment start time, τ. The total attack rate is decreased by decreasing the number of sensitive infections at the treatment start time ($I(\tau)$) and by decreasing the number of susceptibles at the treatment start time ($S(\tau)$) (Equation (4.8)). As κ increases the effect of decreasing $I(\tau)$ becomes more important than decreasing $S(\tau)$, therefore τ^* decreases as κ increases (from left to right the order of the vertical lines is solid, dashed, dot-dashed). **Panel C**: The number of susceptibles as a function of treatment start time. As κ increases $S(\tau^*)$ increases. Also, the points $S = S_{\min}$ are indicated by 'circle' markers and coincide with $S(\tau^*)$. Parameters are $R_I = 1.6$, $R_R = 0.8R_I$ and $\mu_I = \mu_R = \frac{1}{3.3}$.

delay precisely balances these two effects, and this can be readily interpreted graphically. Figure 4.2 (Panel A) shows the final number of susceptibles as a function of treatment start time for different values of κ. As κ increases the importance of the second term in Equation (4.8) increases and correspondingly the number of infecteds at the optimal treatment start time decreases (Figure 4.2, Panel B) while the number of susceptibles at the optimal treatment start time increases (Figure 4.2, Panel C).

4.4.4 Numerical Results for More Detailed Models

Although the analytic expressions presented above are derived using a simple model, numerical simulations suggest that these results hold more generally. To illustrate this, we use numerical simulations to compare the relationship between treatment start time and attack ratio for our simple model and for the more detailed model in [28]. Figure 4.1 (Panel B) depicts the treatment model used in [28] and the corresponding system equations are provided in the Supplemental Information, Section 4.6.5.

To begin, numerical results show that both models exhibit two key features : (i) if the treatment level is constrained to be constant throughout the entire epidemic (as has been the case in most previously published analyses) then there is an intermediate optimal level of treatment; and (ii) if treatment is delayed the appropriate amount, then maximum treatment is always optimal (see Figures 4.4 and 4.5 in Supplemental Information, Section 4.6.5). The fact that delaying treatment can be better than constant treatment in both models explains why [22, 28, 33, 36] observed an intermediate optimal constant treatment level.

A comparison of the panels of Figure 4.3 shows that the general relationship

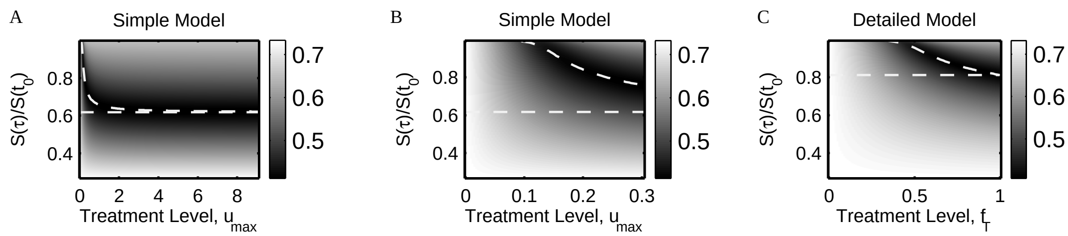

Figure 4.3: **Total attack ratio versus treatment level and number of susceptibles at treatment start time (normalized by $S(t_0)$).** The white dashed curves show the number of susceptibles at the optimal treatment start time provided a resistant epidemic occurs. The white dashed horizontal lines indicates the number of susceptibles at the optimal treatment start time when the maximum possible treatment level is very large (i.e., when $u_{\max}$ is unbounded for the simple model and when $f_T = 1$ for the detailed model). The white dashed horizontal lines in Panels A and B were computed analytically all other curves were computed numerically. It is important to emphasize that this figure was generated by assuming that an outbreak occurs. Indeed, for this specific choice of parameters, Figure 4.6 of the Supplemental Information, Section 4.6.5 illustrates that for $u_{\max} > 0.25$ and $f_T > 0.8$ treating immediately will prevent an outbreak and so treating immediately is optimal. **Panel A**: Figure produced using simple model. **Panel B**: Magnified version of panel A. **Panel C:** Figure produced using detailed model. Parameters are $R_R = 0.9R_I$, $\mu_I = \mu_T = \mu_R = 1/3.3$, $\kappa = 0.0066$, $f_r = 0.002$.

between the treatment level, the attack ratio, and the number of susceptible hosts at the start of treatment, is qualitatively similar for both models. Furthermore, Figure 4.3 (Panel A) shows that for the simple model, if an outbreak occurs, then $S(\tau^*)$ approaches $S_{\min}$ as the treatment level increases (i.e., the solid red curve approaches the horizontal dashed line). In other words, the optimal solution for the case when $u_{\max}$ is very large provides a bound for the case when $u_{\max}$ is finite. A similar statement is true for the more detailed model. In the more detailed model f_T denotes the fraction of infections that are treated and so $f_T = 1$ is analogous to $u_{\max}$ being unbounded (i.e., all infected individuals are treated). Using $S_{\min,c}$ to denote the number of susceptibles at the optimal treatment start time when $f_T = 1$, Figure 4.3 (Panel C) shows that, for the detailed model, $S(\tau^*)$ decreases as the treatment level increases, with $S_{\min,c}$ providing a bound for all other values of f_T.

Both models also demonstrate that it is possible to avoid an outbreak by starting treatment immediately at sufficiently high treatment levels (see Figure 3 in Supplemental Information, Section 4.6.5). Furthermore, if the initial pool of susceptibles is small enough then it is always best to start treatment immediately (see Figure 4.7 in Supplemental Information, Section 4.6.5). For the simple model, 'small enough' is measured relative to the number of susceptibles at the optimal treatment start time when $u_{\max}$ is unbounded. Namely, if $S(t_0) < S_{\min}$ then $\tau^* = t_0$. Similarly for the more detailed model, Figure 4.7 (Panel B) in the Supplemental Information, Section 4.6.5 suggests that if $S(t_0) < S_{\min,c}$ then again it is optimal to start treatment immediately.

Finally, recall that for the simple model, if $R_{IR} > R_R$ then it is optimal to start treatment immediately. Using the biological interpretation of R_{IR}, namely that R_{IR} is the expected total number of infections caused by an individual initially infected with the treatment-sensitive strain, we can define an analogous quantity for the detailed model (see Supplemental Information, Section 4.6.5 for details). Figure 4.8 and Figure 4.9 in Section 4.6.5 shows that the relationship between R_R, R_{IR} and the optimal treatment start time is similar for both the simple and detailed models. This result is particularly interesting because it emphasizes that (i) using a simple model can highlight relationships in more detailed models that may otherwise be undetected and (ii) the quantity R_{IR} is of general importance and not just an artifact of our simple model.

4.5 Discussion

Perhaps the most salient feature of previous analyses is that the best treatment somehow balances the effects of the sensitive and resistant epidemics. This trade-off, which is very clearly explained in the discussion found in [28], is also highlighted in our model. The difference is in the way this trade-off is resolved. In [28] this trade-off is used to explain why an intermediate level of constant treatment is better than a high level of constant treatment. Conversely, we have shown that the optimal treatment balances this trade-off by delaying the onset of treatment instead of capping the maximum level of treatment. By allowing the treatment strategy to vary in time, we have shown that higher levels of treatment are always better, provided they are started at the appropriate time. Furthermore, when the maximum possible treatment level is very large, our analysis shows that the optimal treatment strategy is the one that switches from a sensitive outbreak to a resistant outbreak precisely when the number of susceptible hosts has decreased to the value given by $S_{\min}$.

We have examined two types of models, a simple model that does not explicitly include the dynamics of the treatment class and a more detailed model that does include treatment dynamics. For the simple model we showed, using analytic calculations, that the optimal treatment strategy is "off-on" (i.e., delay treatment for a certain amount of time -possibly no time- and then treat maximally for the remainder of the outbreak). It is important to emphasize that for model (4.1) this "off-on" strategy is indeed optimal in the sense that any other, more complicated, strategy that involves intermediate levels of treatment will not result in a lower attack ratio. Conversely, for the more detailed model we have only shown (via numerical simulations) that an "off-on" strategy can outperform a "constant" strategy. Thus, although it seems likely that the "off-on" strategy is optimal for the more detailed model as well it remains an open problem to explicitly prove that this is indeed the case. Furthermore, although an "off-on" strategy can be better than a "constant" strategy, policy makers still face the ethical dilemma of choosing between the rights of individuals to receive treatment and the benefit to the population of delaying treatment.

Given the ethical problems associated with implementing the optimal treatment strategy when it involves a delay, it is useful to clearly delineate the conditions under which implementing treatment immediately is actually optimal. It is always optimal to start treatment immediately if one of three conditions holds: (i) immediate treatment can prevent an outbreak; (ii) the initial pool of susceptibles is small (i.e., $S(t_0) < S_{\min}$ for the simple model and $S(t_0) < S_{\min,c}$ for the detailed model); or (iii) the rate of de novo resistance is small (i.e., $R_{IR} > R_R$). Each of these three situations have relevance to public health policy.

Condition (i) underscores that our focus has been to minimize the total outbreak size and so in this context preventing an outbreak is the best possible outcome. This outcome however, may not be desirable if there is a high probability of a second epidemic occurring at a later time. For example, although treating immediately at a very high level may minimize the total attack ratio, this will result in $S(t_f) > S_R > S_I$. As a result, the population will be susceptible to future sensitive and resistant outbreaks. Alternatively, if a treatment strategy is chosen to ensure that a resistant epidemic occurs, then the final number of susceptibles will be small enough to prevent future resistant outbreaks (and the severity of a possible future sensitive outbreak will be significantly decreased). These issues are discussed in [8] in the context of a single-strain epidemic. Interestingly, if delay is optimal then, up to a point, increasing

$u_{\max}$ will increase the optimal amount of delay. However, if $u_{\max}$ can be increased sufficiently then treating immediately will effectively prevent the outbreak, and this then becomes the optimal strategy. If this can occur then policy makers will have to balance the benefits of treating immediately (and therefore preventing the current outbreak) with the benefits of delaying treatment (and perhaps minimizing the effects of a second outbreak).

Condition (ii) highlights that the optimal treatment strategy depends on the size of the target community. Essentially, if a community is small enough then treatment should start immediately. This observation leads to an interesting hypothesis for combined vaccination-treatment policies. Namely, since vaccination decreases the susceptible population, low vaccination levels make it more likely that treatment should be delayed. This proposed relationship is surprising since, naively, it seems reasonable to expect that if fewer individuals are vaccinated then it would be better to start treatment early in order to reduce spread. Nevertheless our results suggest that this intuition is, in fact, incorrect.

Condition (iii) emphasizes that if the probability of receiving treatment is low, then it is more likely that treatment should start immediately. More concretely, by expressing R_{IR} as a linear combination of the two reproduction numbers R_I and R_R we can rewrite $R_{IR} > R_R$ as

$$\frac{R_I}{R_R} > \frac{1 - \kappa\chi}{1 - \chi}.$$

Therefore, if $\frac{R_I}{R_R} > \frac{1-\kappa\chi}{1-\chi}$ it is optimal to start treatment immediately and if $\frac{R_I}{R_R} \leq \frac{1-\kappa\chi}{1-\chi}$ then it may be best to delay treatment. For any specific values of R_I, R_R and κ we see that as the probability of treating an infected individual, χ, decreases, it becomes more likely that starting treatment immediately is best. Intuitively this guideline makes sense since if treatment is unlikely to reach many infected individuals then it makes sense to start administering treatment as soon as possible in order to reach more individuals.

Since multiple intervention measures are often used to control a disease, an important next step is to consider a model that includes other intervention measures, in addition to treatment [16, 17, 20, 29]. Some models similar to our model that have included multi-intervention strategies are the treatment-vaccination models in [18, 36] and the treatment-prophylaxis model in [28]. Interestingly, a number of studies that consider multi-intervention strategies also exhibit optimal constant intervention levels. This suggests that considering multi-intervention strategies that vary in time may be very informative as well.

4.6 Supplemental Information

4.6.1 Construction of the Simplified Model

The model depicted in Figure 4.1 is described by the following system of equations:

$$\begin{aligned}
\dot{S} &= -(\beta_I I + \beta_T T + \beta_R R)S, \\
\dot{I} &= (\beta_I I + \beta_T T)S - (\mu_I + u)I, \\
\dot{T} &= uI - (\mu_T + \nu)T, \\
\dot{R} &= \beta_R RS + \nu T - \mu_R R. \qquad (4.9)
\end{aligned}$$

To facilitate mathematical analysis we make two assumptions. The first assumption is that treatment is perfectly effective, meaning that essentially no transmission occurs from treated individuals. Under this assumption model (4.9) becomes,

$$\begin{aligned}
\dot{S} &= -(\beta_I I + \beta_R R)S, \\
\dot{I} &= \beta_I I S - (\mu_I + u)I, \\
\dot{T} &= uI - (\mu_T + \nu)T, \\
\dot{R} &= \beta_R R S + \nu T - \mu_R R.
\end{aligned} \tag{4.10}$$

The second assumption is that the dynamics of the treated class are very fast relative to the other classes. With this assumption, $\dot{T} \sim 0$ for all t. In this case, $T(t) \sim \frac{u(t)I(t)}{\mu_T+\nu}$ and so model (4.10) can be simplified to

$$\begin{aligned}
\dot{S} &= -(\beta_I I + \beta_R R)S, \\
\dot{I} &= \beta_I I S - (\mu_I + u)I, \\
\dot{R} &= \beta_R R S + \kappa u I - \mu_R R,
\end{aligned} \tag{4.11}$$

where $\kappa = \frac{\nu}{\nu+\mu_T}$. In other words, with these assumptions, we are effectively assuming that treatment causes one of two things to happen instantaneously: (i) the treated individual is removed from the infectious population with probability $(1-\kappa)$ or (ii) the treated individual develops resistance with probability κ.

For model (4.11)

$$\int_{t_0}^{t_f} \beta_I I S + \beta_R R S \mathrm{d}t = S(t_0) - S(t_f)$$

and therefore minimizing the total attack ratio is equivalent to minimizing the total outbreak size which is equivalent to maximizing the final number of susceptibles (assuming that $S(t_0)$ is fixed).

An interesting observation from model (4.11) is that the number of susceptibles is a strictly decreasing function of time. This suggests that either "the current number of susceptibles" or "the current time" can be used to track the current state of an outbreak. So for example if we introduce a new variable $y = -S$ then $\frac{\mathrm{d}y}{\mathrm{d}t} = (\beta_I I + \beta_R R)S$ and model (4.11) can be written equivalently as:

$$\begin{aligned}
\dot{S} &= -1, \\
\dot{I} &= \frac{\beta_I IS - (\mu_I + u)I}{(\beta_I I + \beta_R R)S}, \\
\dot{R} &= \frac{\beta_R RS + \kappa u I - \mu_R R}{(\beta_I I + \beta_R R)S},
\end{aligned} \tag{4.12}$$

where a "dot" means differentiation with respect to y. In Section 4.6.4 we will use the equations in (4.12) instead of the equations in (4.11) to solve for the optimal treatment strategy. Although both sets of equations are equivalent, when we apply Pontryagin's Maximum Principle the equations in (4.12) lead to simpler expressions and thus we are able to proceed further analytically. Although y is not in fact a "time" variable we can treat it as a "time" variable because it is a strictly increasing function of time. For simplicity, in Section 4.6.4 we will refer to y as time and use the symbol t instead of y.

4.6.2 Definition of the End of the Outbreak

The end of the epidemic in models such as system (4.11) is not well defined, because the number of infected individuals approaches zero asymptotically as time increases to infinity. As a result, it is not immediately clear how best to choose t_f. The three most natural choices are

(i) $t_f = \infty$,

(ii) $t_f = t_{\max}$, where $t_{\max}$ is a sufficiently large constant, and

(iii) $t_f = \inf\{t \mid I(t) + R(t) = r\}$, where r is some constant chosen to indicate the end of the epidemic.

A desirable property of the definition of "end of an outbreak" is that it is consistent across analysis of different models and varying parameter values. A problem with method (i) is that we can not use it for numerical analysis because numerical simulations are, by necessity, finite in duration. A problem with method (ii) is that it is not clear how large $t_{\max}$ needs to be in order to coincide with a reasonable interpretation of "the end of the epidemic". Furthermore, how large $t_{\max}$ needs to be will be different for different treatment strategies and different parameter values. On the other hand, method (iii) is consistent across analysis of different models and parameter values and additionally agrees with intuition since it says that an outbreak is over once the number of infectious individuals (of all types) is less than a predefined small number.

4.6.3 Analysis for Situation 1

Optimal Treatment Strategy when $u_{\max} \to \infty$

In the limit that $u_{\max} \to \infty$, model (4.11) is essentially the combination of the systems,

$$\begin{aligned} \dot{S} &= -\beta_I I S, \\ \dot{I} &= \beta_I I S - \mu_I I, \end{aligned} \tag{4.13}$$

with $(S(t_0), I(t_0)) = (S_0, I_0)$ and

$$\begin{aligned} \dot{S} &= -\beta_R R S, \\ \dot{R} &= \beta_R R S - \mu_R R, \end{aligned} \tag{4.14}$$

with $(S(\tau), R(\tau)) = (S(\tau), \kappa I(\tau))$. In short, when $u_{\max} \to \infty$, determining the optimal treatment strategy becomes the problem of deciding when to switch from model (4.13) to model (4.14).

Integrating and rearranging the Equations in (4.13) gives

$$S_I \ln S(\tau) - S(\tau) = S_I \ln S_0 - S_0 - I_0 + I(\tau). \tag{4.15}$$

Integrating and rearranging the Equations in (4.14) gives

$$S_R \ln S(t_f) - S(t_f) = S_R \ln S(\tau) - S(\tau) - R(\tau) + R(t_f). \tag{4.16}$$

Noting that $R(\tau) = \kappa I(\tau)$ and substituting Equation (4.15) into Equation (4.16) gives

$$S_R \ln S(t_f) - S(t_f) = (S_R - \kappa S_I) \ln S(\tau) - (1-\kappa) S(\tau) + C, \tag{4.17}$$

where $C = \kappa(S_I \ln S_0 - S_0 - I_0) + R(t_f)$.

Situation 1.A: If a resistant outbreak occurs then the number of resistant individuals will increase until $S = S_R$. Therefore, if a treatment start time results in $S(t_f) > S_R$ then starting treatment at this time prevents the resistant outbreak from occurring. If a resistant outbreak is avoided then this implies that starting treatment immediately ends the outbreak ($\tau = t_f$). This implies that, $\kappa I(\tau) = \kappa I(t_f) \leq r$ and $S(\tau) = S(t_f)$. Therefore, since we want to maximize $S(t_f)$ if treatment can prevent the resistant outbreak then it is best to treat immediately. This corresponds to Situation 1.A of the main text.

Situation 1.B: If on the other hand treating immediately cannot prevent the resistant outbreak then starting treatment at any time will results in $S(t_f) < S_R$. In this case the treatment start time that maximizes the $S(t_f)$ also maximizes the right-hand side of Equation (4.17). (This is because $S_R \ln S(t_f) - S(t_f)$ is a convex down function of $S(t_f)$ with a maximum at $S(t_f) = S_R$.)

There are two cases:

(i) $\kappa < 1$: In this case Equation (4.17) simplifies to

$$S_R \ln S(t_f) - S(t_f) = (1-\kappa)(S_{\min} \ln S(\tau) - S(\tau)) + C,$$

where C is a constant (since $R(t_f) = r$). Therefore the optimal treatment start time τ^* is defined by

$$S(\tau^*) = \min\{S_0, S_{\min}\}.$$

This corresponds to Situation 1.B of the main text.

(ii) $\kappa = 1$: In this case Equation (4.17) becomes,

$$S_R \ln S(t_f) - S(t_f) = (S_R - S_I) \ln S(\tau) + C.$$

Therefore,

(a) if $S_R - S_I > 0$ then treatment should start immediately,

(b) if $S_R - S_I = 0$ then treatment can start at any time, and

(c) if $S_R - S_I < 0$ then it is optimal to never treat.

Derivation of Equation (4.8) in Main Text

If A denotes the total attack ratio, A_I denotes the sensitive attack ratio and A_R denotes the resistant attack ratio then,

$$\begin{aligned}
\frac{\mathrm{d}A}{\mathrm{d}\tau} &= \frac{\mathrm{d}A_I}{\mathrm{d}\tau} + \frac{\mathrm{d}A_R}{\mathrm{d}\tau}, \\
&= \frac{\mathrm{d}}{\mathrm{d}\tau}\frac{S_0 - S(\tau)}{S_0} + \frac{\mathrm{d}}{\mathrm{d}\tau}\frac{S(\tau) - S(t_f)}{S_0}, \\
&= -\frac{1}{S_0}\left[\frac{\mathrm{d}S(\tau)}{\mathrm{d}\tau} - \frac{\mathrm{d}S(\tau)}{\mathrm{d}\tau} + \frac{\mathrm{d}S(t_f)}{\mathrm{d}\tau}\right], \\
&= -\frac{1}{S_0}\frac{\partial S(\tau)}{\partial \tau}\left[\frac{\partial S(\tau)}{\partial S(\tau)} - (\frac{\partial S(\tau)}{\partial S(\tau)} - \frac{\partial S(t_f)}{\partial S(\tau)}) + \frac{\partial S(t_f)}{\partial I_\tau}\frac{\partial I_\tau}{\partial S(\tau)}\right], \qquad (4.18) \\
&= \frac{\partial S(\tau)}{\partial \tau}\left[\frac{\partial A_I}{\partial S(\tau)} + \frac{\partial A_R}{\partial S(\tau)} + \frac{\partial A_R}{\partial I_\tau}\frac{\partial I_\tau}{\partial S(\tau)}\right]. \qquad (4.19)
\end{aligned}$$

The three effects of treatment are very clearly delineated by Equations (4.18) and (4.19). First, treatment is beneficial because it suppresses the spread of the sensitive strain. Correspondingly the first term is positive and indicates that increasing delay increases the sensitive attack ratio. If the aim is to suppress the sensitive strain then it is best to start treatment immediately. Second, treatment is detrimental because suppressing the sensitive strain leaves a larger pool of susceptible hosts for the resistant strain. Correspondingly the second term is negative and indicates that increasing delay decreases the resistant attack ratio. If the aim is to decrease the susceptible pool for the resistant outbreak then it is best to delay treatment. Third, treatment is detrimental because it leads to the de novo appearance of resistant infections. Correspondingly the third term is positive and indicates that increasing delay increases the resistant attack ratio. If the aim is to decrease the amount of de novo resistance then it is best to treat immediately.

To derive Equation (4.8) in the main text it only remains to compute $\frac{\partial S(t_f)}{\partial S(\tau)}$ and $\frac{\partial S(t_f)}{\partial I(\tau)}\frac{\mathrm{d}I(\tau)}{\mathrm{d}S(\tau)}$ and substitute these expressions into Equation (4.18). Let $G = S_R \ln S(t_f) - S(t_f)$, then by Equation (4.16) we have,

$$G = S_R \ln S(t_f) - S(t_f) = S_R \ln S(\tau) - S(\tau) - \kappa I(\tau) + R(t_f). \tag{4.20}$$

Therefore,

$$\begin{aligned} \frac{\partial G}{\partial S(\tau)} &= \frac{\partial G}{\partial S(t_f)}\frac{\partial S(t_f)}{\partial S(\tau)}, \\ \frac{S_R}{S(\tau)} - 1 &= (\frac{S_R}{S(t_f)} - 1)\frac{\partial S(t_f)}{\partial S(\tau)}. \end{aligned} \tag{4.21}$$

Rearranging Equation (4.21), we have $\frac{\partial S(t_f)}{\partial S(\tau)} = \frac{\frac{S_R}{S(\tau)}-1}{\frac{S_R}{S(t_f)}-1}$. Similarly,

$$\begin{aligned} \frac{\partial G}{\partial I(\tau)} &= \frac{\partial G}{\partial S(t_f)}\frac{\partial S(t_f)}{\partial I(\tau)}, \\ -\kappa &= (\frac{S_R}{S(t_f)} - 1)\frac{\partial S(t_f)}{\partial I(\tau)}. \end{aligned} \tag{4.22}$$

Rearranging Equation (4.22), we have $\frac{\partial S(t_f)}{\partial I(\tau)} = \frac{-\kappa}{\frac{S_R}{S(t_f)}-1}$. Finally, by Equation (4.15) we have,

$$\frac{\partial I(\tau)}{\partial S(\tau)} = (\frac{S_I}{S(\tau)} - 1). \tag{4.23}$$

Substituting these expressions into Equation (4.18) gives

$$\frac{\mathrm{d}A}{\mathrm{d}\tau} = \alpha[S_R - S(\tau)] - \alpha\kappa[S_I - S(\tau)], \tag{4.24}$$

where $\alpha = \frac{S(t_f)\beta_I I(\tau)}{S(t_0)(S_R - S(t_f))} > 0$. Equation (4.24) summarizes the three effects of treatment in terms of (i) the effect of changing the size of the susceptible pool for the resistant strain (i.e., the first two terms in Equation (4.19) combine to form the first term Equation (4.24)) and (ii) the effect of changing the amount of de novo resistance (i.e., the third term in Equation (4.19) becomes the second term in Equation (4.24)).

4.6.4 Analysis for Situation 2

Basic Form of the Optimal Treatment Strategy

In this section we use Pontryagin's Maximum Principle (PMP) to solve the following problem:

Problem 4.1. *Determine the treatment strategy u for model* (4.12) *that minimizes the total attack ratio*

$$\frac{S(t_0) - S(t_f)}{S(t_0)}$$

subject to $S(t_0) = S_0$, $I(t_0) = I_0 > 1$, $R(t_0) = 0$, $t_f = \min\{t|I(t) + R(t) \leq 1\}$ and $u(t) \in [0, u_{\max}]$ for all $t \in [t_0, t_f]$. Notice that since $S(t_0)$ is fixed, minimizing the total attack ratio is equivalent to maximizing the final number of susceptibles $S(t_f)$.

Our discussion of PMP is non-technical and limited to the context of Problem 4.1. For example, among other things, we omit details regarding the smoothness properties of certain functions. Additionally we do not show how to prove that an optimal control actually exists (although this can be done with a straightforward application of an existance theorem found in [19]). For a more comprehensive discussion of PMP see [13, 26, 27, 35].

We begin by defining a number of functions. First, we define a function H (called the Hamiltonian):

$$H = \lambda_S \dot{S} + \lambda_I \dot{I} + \lambda_R \dot{R}.$$

The definition of the Hamiltonian involves the (still undefined) functions λ_S, λ_I, and λ_R (called the adjoint functions). The adjoint functions are defined by the following set of differential equations

$$\begin{aligned} \dot{\lambda}_S &= -\frac{\partial H}{\partial S}, \\ \dot{\lambda}_I &= -\frac{\partial H}{\partial I}, \\ \dot{\lambda}_R &= -\frac{\partial H}{\partial R}, \end{aligned} \tag{4.25}$$

and endpoint conditions: $(\lambda_S(t_f), \lambda_I(t_f), \lambda_R(t_f)) = (\lambda_0, q, q)$ where λ_0 and q are constants, $q \leq 0$ and λ_0 is equal to either 0 or 1. Furthermore, the vector

$$\vec{\lambda} = (\lambda_0, \lambda_S(t), \lambda_I(t), \lambda_R(t))$$

is nonzero for all $t \in [t_0, t_f]$.

An essential element of any control problem is that by changing the input (or control) we are able to change the output. So in the context of model (4.12) by changing the treatment strategy u we are able to change the states $\vec{x} = (S, I, R)$

and thus change the course of the outbreak. Similarly, the Hamiltonian and the adjoint functions are also affected by the treatment strategy (i.e., they depend on u). Suppose an optimal treatment strategy u^* exists. Then let $\vec{x}^* = (S^*, I^*, R^*)$, $\vec{\lambda}^* = (\lambda_0, \lambda_S^*, \lambda_I^*, \lambda_R^*)$ and $H^* = H(\vec{\lambda}^*, \vec{x}^*, u^*)$ be the states, the adjoint functions and the Hamiltonian that result from using the optimal treatment strategy.

Up to this point we have just defined a number of functions that satisfy certain properties. The key result of the PMP is that if an optimal treatment strategy exists then that treatment strategy maximizes the Hamiltonain. Furthermore, PMP states that because the specific value of t_f is not fixed, the Hamiltonian is equal to zero. More concretely,

$$0 = H(\vec{\lambda}^*(t), \vec{x}^*(t), u^*(t)) \geq \sup_{z\in[0,u_{\max}]} H(\vec{\lambda}^*(t), \vec{x}^*(t), z) \tag{4.26}$$

for almost all $t \in [t_0, t_f]$.

We will now use the equations for model (4.12) to examine the consequences of relation (4.26) and see how it can be used to characterize the form of the optimal treatment strategy.

Using the equations for model (4.12) the Hamiltian can be expressed as

$$H = -\lambda_S + \frac{\beta_I S I \lambda_I + \beta_R S R \lambda_R - \lambda_I \mu_I I - \lambda_R \mu_R R + (\lambda_R \kappa - \lambda_I) u I}{(\beta_I I + \beta_R R) S} = 0, \tag{4.27}$$

where the adjoint vector $\vec{\lambda} = (\lambda_0, \lambda_S, \lambda_I, \lambda_R)$ satisfies the following set of differential equations

$$\begin{aligned}
\dot{\lambda}_0 &= 0, \\
\dot{\lambda}_S &= \frac{-\lambda_I \mu_I I - \lambda_R \mu_R R + (\lambda_R \kappa - \lambda_I) u I}{(\beta_I I + \beta_R R) S^2}, \\
\dot{\lambda}_R &= \left((S - S_R)(\lambda_I - \lambda_R) + \lambda_I (S_R - S_I) + (\lambda_R \kappa - \lambda_I) \frac{u}{\beta_I} \right) \frac{\beta_I \beta_R I}{(\beta_I I + \beta_R R)^2 S}, \qquad (4.28) \\
\dot{\lambda}_I &= -\dot{\lambda}_R \frac{R}{I}. \qquad (4.29)
\end{aligned}$$

Furthermore, notice that since $H = 0$ and $\vec{\lambda}(t_f) \neq 0$ it must be that q is strictly negative and so the endpoint conditions for the adjoint functions are

$$(\lambda_0, \lambda_S(t_f), \lambda_I(t_f), \lambda_R(t_f)) = (\lambda_0, \lambda_0, q, q)$$

where λ_0 is either 0 or 1 and $q < 0$.

Now because the Hamiltonian is a linear function of u (see Equation (4.27)) and the optimal treatment strategy maximizes the Hamiltonian (see relation (4.26)) we see that the optimal treatment strategy is determined by the sign of $\frac{\partial H}{\partial u}$. Namely if $\frac{\partial H}{\partial u} > 0$ then larger values of u result in larger values of H and hence $u = u_{\max}$ maximizes H. On the other hand if $\frac{\partial H}{\partial u} < 0$ then smaller values of u result in larger values of H and hence $u = 0$ maximizes H. Now since, $\frac{\partial H}{\partial u} = \frac{(\lambda_R \kappa - \lambda_I) I}{(\beta_I I + \beta_R R) S}$ the sign of $\frac{\partial H}{\partial u}$ is the same as the sign of $(\lambda_R \kappa - \lambda_I)$. Therefore, the optimal treatment strategy satisfies

$$u^* = \begin{cases} u_{\max} & \text{if } \lambda_R \kappa - \lambda_I > 0 \\ ? & \text{if } \lambda_R \kappa - \lambda_I = 0 \\ 0 & \text{if } \lambda_R \kappa - \lambda_I < 0. \end{cases}$$

We will call $\lambda_R\kappa - \lambda_I$ the "switch function" and denote it by $\text{SF} = (\lambda_R\kappa - \lambda_I)$. If $\text{SF} = 0$ for an interval of time $[t_1, t_2]$ then relation (4.26) does not charactersize the optimal treatment strategy. In this case the optimal treatment strategy is called "singular" on the interval $[t_1, t_2]$ and more work is required to determine what u^* should be on $[t_1, t_2]$. For Problem (4.1) the optimal control is singular only in very special circumstances and these are discussed in Section 4.6.6 of this Supplementary Information. For the rest of this discussion we will always assume that the optimal treatment strategy is non-singular.

In Section 4.6.6 we also show that SF is a monotonic function of time. This means that the value of u^* can switch at most once. Additionally, $\text{SF}(t_f) = \lambda_R(t_f)\kappa - \lambda_I(t_f) = q(\kappa - 1) \geq 0$ and therefore if the value of u^* does switch then it must switch from $u = 0$ to $u = u_{\max}$. Using this information, we conclude that the optimal treatment strategy is either

(i) to treat maximally for the entire epidemic ($u^* \equiv u_{max}$),

(ii) to delay the start of treatment and then to treat maximally for the remainder of the epidemic, or

(iii) to never treat ($u^* \equiv 0$).

The problem of finding the optimal treatment policy now reduces to the problem of finding the optimal treatment start time τ^*. Note that if $\tau^* = t_0$ (that is, the optimal treatment start time is the start of the epidemic) then this corresponds to situation (i). If $\tau^* \in (t_0, t_f)$ then this corresponds to situation (ii). Finally, if $\tau^* = t_f$ (that is, the optimal treatment start time is the end of the epidemic) then this corresponds to situation (iii). Notice that situation (iii) is possible only if $\text{SF}(t_f) = q(\kappa - 1) = 0$ and therefore, the only time that it is optimal to never treat is when $\kappa = 1$.

The Optimal Amount of Delay when $\kappa < 1$

The optimal amount of delay when $\kappa < 1$ is characterized by the relation

$$S(\tau^*) \geq \min(S(t_0), S_{\min}).$$

This relation means that treatment should be delayed no longer than the time it takes for the number of susceptibles to decrease to $S_{\min}$. Furthermore, if the initial number of susceptibles is less than $S_{\min}$ then treatment should begin immediately.

From the previous section we know that if $\kappa < 1$ then the optimal treatment policy is to either (i) start treatment immediately and treat with maximum effort for the entire outbreak or (ii) delay the start of treatment and then treat with maximal effort for the entire outbreak. If policy (i) is optimal then $S(t_0) = S(\tau^*)$ and so clearly $S(\tau^*) \geq \min(S(t_0), S_{\min})$.

Suppose policy (ii) is optimal. In this case the sign of SF must switch from negative to positive at $t = \tau$ and hence SF is an increasing function of time. By Equation (4.29), we have that $\dot{\mathrm{SF}} = (\dot{\lambda}_R\kappa - \dot{\lambda}_I) = \dot{\lambda}_R(\kappa + \frac{R}{I})$ and therefore SF is an increasing function of time only if $\dot{\lambda}_R \geq 0$. This implies, by Equation (4.28), that

$$(S - S_R)(\lambda_I - \lambda_R) + \lambda_I(S_R - S_I) + (\lambda_R\kappa - \lambda_I)\frac{u}{\beta_I} \geq 0. \tag{4.30}$$

Evaluating Equation (4.30) at $t = \tau^*$ and using the fact that $\mathrm{SF}(\tau^*) = \lambda_R(\tau^*)\kappa -$

$\lambda_I(\tau^*) = 0$ gives,

$$\lambda_R(\tau^*)[(S(\tau^*) - S_R)(\kappa - 1) + \kappa(S_R - S_I)] \geq 0$$

and therefore

$$(S(\tau^*) - S_R)(\kappa - 1) + \kappa(S_R - S_I) \leq 0 \tag{4.31}$$

since $\lambda_R < 0$ (see Remark 4.1). Rearranging Equation (4.31) gives,

$$S(\tau^*) \geq S_R + \frac{\kappa}{1-\kappa}(S_R - S_I) = S_{\min}.$$

So clearly, if policy (ii) is optimal then $S(\tau^*) \leq \min(S(t_0), S_{\min})$.

Remark 4.1. We know that $\lambda_R < 0$ because $\dot{\lambda}_R \geq 0$ and $\lambda_R(t_f) = q < 0$.

Optimal Treatment Strategy when $\kappa = 1$

If $\kappa = 1$ then $\mathrm{SF}(t_f) = \lambda_R(t_f) - \lambda_I(t_f) = 0$. Furthermore we know that SF and λ_R are monotonic functions of time (see Section 4.6.6). By Equation (4.28) we have,

$$\dot{\mathrm{SF}}(t_f) = \dot{\lambda}_R(t_f)\left[\kappa + \frac{R(t_f)}{I(t_f)}\right] = (S_R - S_I)q\frac{\beta_I\beta_R I(t_f)}{(\beta_I I(t_f) + \beta_R R(t_f))^2 S(t_f)}\left[\kappa + \frac{R(t_f)}{I(t_f)}\right],$$

where $q < 0$. Therefore the sign of $\dot{\mathrm{SF}}$ is equal to the sign of $q(S_R - S_I)$. Since $\mathrm{SF}(t_f) = 0$ this implies that $u^* \equiv 0$ if $(S_R - S_I) > 0$ and $u^* \equiv u_{\max}$ if $(S_R - S_I) < 0$. Finally, in Section 4.6.6 we show that in the case that $(S_R - S_I) = 0$ the optimal treatment strategy is singular.

Optimal treatment strategy when $\kappa = 0$

If $\kappa = 0$ then $R(t) \equiv 0$ which implies that $\dot{\mathrm{SF}} = \dot{\lambda}_R(\kappa + \frac{R}{I}) = 0$. This means that SF is constant. Furthermore, $\mathrm{SF}(t_f) = \lambda_R(t_f)\kappa - \lambda_I(t_f) = -\lambda_I(t_f) = -q > 0$. Therefore, $\mathrm{SF}(t) > 0$ for all t which implies that $u^* \equiv u_{\max}$.

If $R_{IR} > R_R$ then Start Treatment Immediately

By Equation (4.29), $\dot{\mathrm{SF}} = (\dot{\lambda}_R\kappa - \dot{\lambda}_I) = \dot{\lambda}_R(\kappa + \frac{R}{I})$. Furthermore, since the switch function SF is monotonic either

(i) $\dot{\mathrm{SF}} = \dot{\lambda}_R(\kappa + \frac{R}{I}) \leq 0$ and hence $\dot{\lambda}_R(t) \leq 0$ for all t.

(ii) $\dot{\mathrm{SF}} = \dot{\lambda}_R(\kappa + \frac{R}{I}) \geq 0$ and hence $\dot{\lambda}_R(t) \geq 0$ for all t, or

Suppose that (i) holds. Then, since $\mathrm{SF}(t_f) \geq 0$, it is optimal to start treatment right away (that is, $u^* \equiv u_{\max}$).

Now, rearranging the relation $R_{IR} > R_R$ gives,

$$0 < (S_R - S_I + (\kappa - 1)\frac{u_{\max}}{\beta_I}). \tag{4.32}$$

Multiplying both sides of (4.32) by $\lambda_R(t_f)$ gives,

$$\begin{aligned} 0 &> \lambda_R(t_f)(S_R - S_I + (\kappa - 1)\frac{u_{\max}}{\beta_I}), \\ 0 &> \dot{\lambda}_R(t_f), \end{aligned} \tag{4.33}$$

where (4.33) uses the fact that $\lambda_R(t_f) = \lambda_I(t_f) = q < 0$. Therefore, $R_{IR} > R_R$ implies that $\dot{\lambda}_R(t) < 0$ for all t which implies that it is optimal to start treatment immediately.

4.6.5 Detailed Model

Definition of Detailed Model

In the detailed model used in the main text an individual can be susceptible (S), infected with a sensitive strain (I), infected with a sensitive strain and treated (T), or infected with a resistant strain (R). Susceptible individuals become infected with the sensitive strain through contact with individuals in the I and T compartments (with rate constants β_I and β_T respectively), and become infected with the resistant strain through contact with individuals in the R compartment (with rate constant β_R). A fraction f_T of individuals infected with the sensitive strain are immediately treated and a fraction f_r of these treated individuals immediately develop resistance. Finally, individuals in each of the three infected classes can leave the system, through either death or recovery with immunity, at constant per capita rates μ_I, μ_T, and μ_R.

The detailed model is described by the following set of equations:

$$
\begin{aligned}
\dot{S} &= -\beta_I SI - \beta_T ST - \beta_R SR, \\
\dot{I} &= (1 - f_T)(\beta_I SI + \beta_T ST) - \mu_I I, \\
\dot{T} &= f_T(1 - f_R)(\beta_I SI + \beta_T ST) - \mu_T T, \\
\dot{R} &= f_T f_R(\beta_I SI + \beta_T ST) + \beta_R SR - \mu_R R.
\end{aligned}
$$

Results involving R_{IR}

Recall that for the simple model, if $R_{IR} > R_R$ then it is always optimal to start treatment immediately. Numerical analysis suggests that this is also true for the more detailed model. The quantity R_{IR} is defined for the detailed model by

$$R_{IR} = \left(\frac{\beta_I}{\mu_I}(1-f_T) + f_T(1-f_R)\frac{\beta_T}{\mu_T} + f_T f_R \frac{\beta_R}{\mu_R}\right) S(t_0).$$

The first term represents the expected number of secondary infections generated by an untreated individual, multiplied by the probability of not being treated; the second term is the expected number of secondary infections generated by a successfully treated individual multiplied by the probability of being successfully treated; and the third term is the expected number of secondary infections generated by a treated individual who develops resistance multiplied by the probability of developing resistance.

Figure 4.4 and Figure 4.5 show that the simple model and the detailed model exhibit a similar relationship between delay and R_{IR}. The dotted curves show the treatment level at which $R_{IR} = R_R$ (as a function of probability of developing resistance). The dashed and solid curves show the treatment level above which some amount of delay is optimal. Since the dashed and solid curves always lie on or above the dotted curves, this suggests that if $R_{IR} > R_R$ then it is not optimal delay.

It is also interesting to note that the dashed and solid curves deviate substantially from the corresponding dotted curve only for very small or large values of κ (and f_R). This is because when $\kappa = 0$ ($f_R = 0$) both models become the standard treatment model (without resistance) and so treating immediately is optimal. Also, when κ (and f_R) become very large most treated individuals develop resistance and since $R_R < R_I$ it is preferable to transfer individuals to the resistant class instead of keeping them in the sensitive class. In other words, for large κ, treatment is predominantly a way to transfer from sensitive to resistant infections instead of a way of removing infections.

Simulation Results Comparing Simple and Detailed Model

In this section we present the figures that substantiate the claims made in Section 4.4.4.

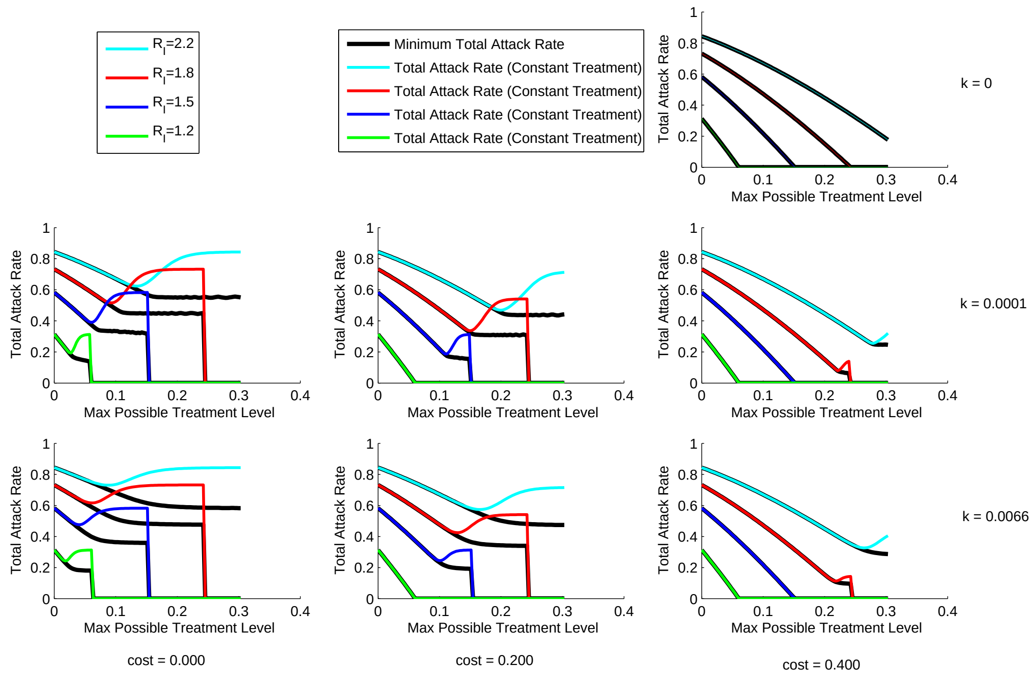

Figure 4.4: **Comparison of minimum total attack ratio and total attack ratio for constant treatment for Simple Model.** Black curves show the minimum total attack ratio for the value of R_I indicated by the colored curve on top of the black curve. The minimum total attack ratio was computed using the optimal treatment start time. The colored curves show the total attack ratio for constant treatment. This figure shows how attack ratio varies with respect to R_I, fitness cost $(1 - \frac{R_R}{R_I})$, and probability of de novo resistance $(\kappa,\ f_R)$. The black curves are monitonically decreasing revealing that the highest possible treatment level of always best provided it is used at the appropriate time. The coloured curves are initially decreasing and then increase again, revealing an optimal intermediate treatment level if treatment is constrained to remain constant throughout the entire outbreak. Additionally we see that (i) for a fixed treatment level, a larger fitness cost leads to a smaller attack ratio and (ii) a greater probability of de novo resistance leads to a larger attack ratio. Also, decreasing R_I decreases the attack ratio. Figure produced using simple model (Equations 1 of main text). Parameters are $R_R = 0.9R_I$, $\mu_I = \mu_R = \frac{1}{3.3}$.

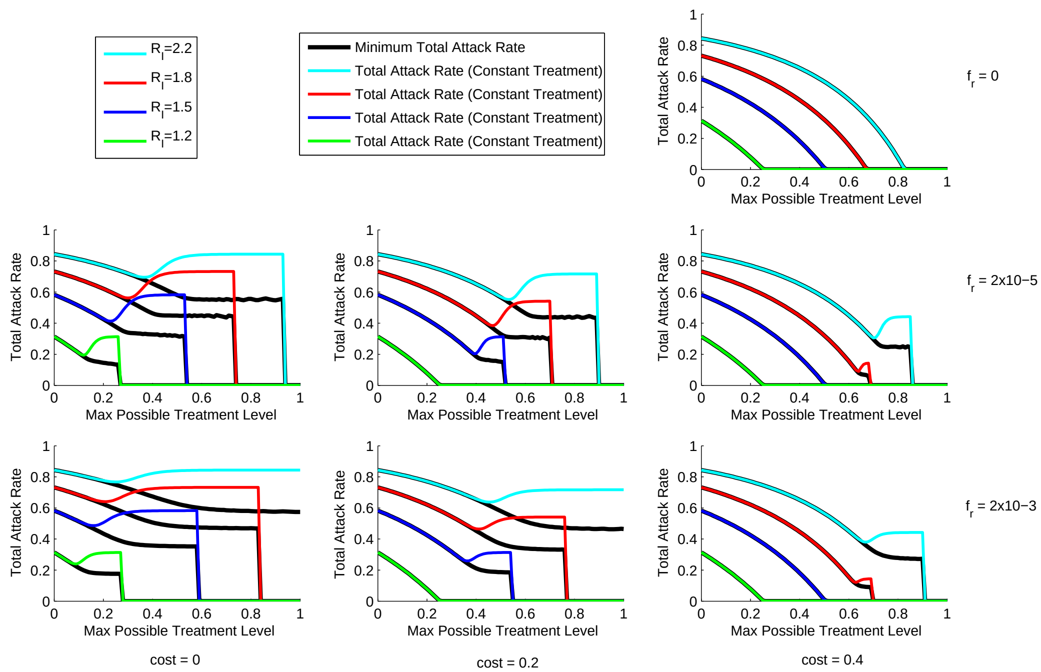

Figure 4.5: **Comparison of minimum total attack ratio and total attack ratio for constant treatment for Detailed Model.** Figure was produced using the detailed model and is analogous to Figure 4.4. Black curves show the minimum total attack ratio for the value of R_I indicated by the colored curve on top of the black curve. The minimum total attack ratio was computed using the optimal treatment start time. The colored curves show the total attack ratio for constant treatment. The black curves are monitonically decreasing revealing that the highest possible treatment level of always best provided it is used at the appropriate time. The coloured curves are initially decreasing and then increase again, revealing an optimal intermediate treatment level if treatment is constrained to remain constant throughout the entire outbreak. Parameters are $R_R = 0.9R_I$, $R_T = 0.34R_I$, $\mu_I = \mu_T = \mu_R = \frac{1}{3.3}$.

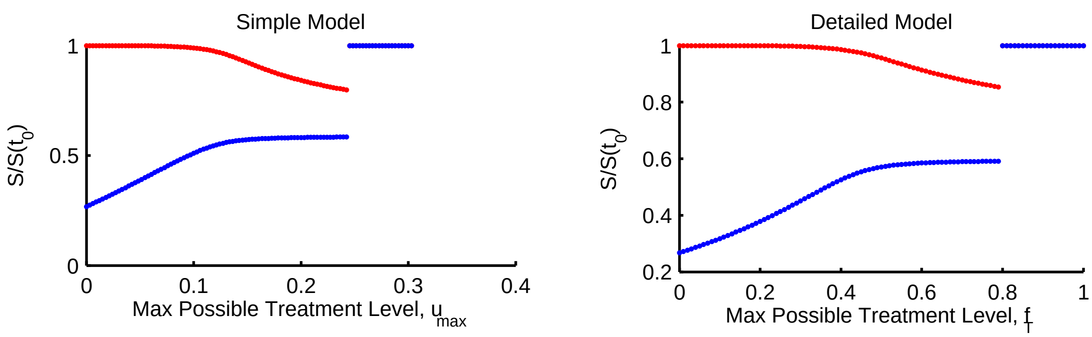

Figure 4.6: **High treatment levels can prevent an outbreak.** Starting treatment immediately can sometimes prevent an outbreak. **Panel A:** Figure produced using the simple model. The red curve indicates the number of susceptibles at the optimal treatment start time as a function of treatment level. The number of susceptibles at the optimal treatment start time decreases (i.e., the optimal treatment start time increases) until $u_{\max} = 0.25$. For $u_{\max} > 0.25$ it is optimal to start treatment immediately (for $u_{\max} > 0.25$ the red line is obscured by the blue line).The blue curve indicates the final number of susceptibles when treatment starts at the optimal time. For very high treatment levels the number of susceptibles at the optimal switch time is $S(t_0)$ and the corresponding final number of susceptibles is just slightly less than $S(t_0)$. In other words, for very high treatment levels it is possible to prevent an outbreak by starting treatment immediately (scenario 1.A$'$). **Panel B:** Figure produced using the detailed model. Figure is similar to that in Panel A. Parameters are $R_R = 0.9R_I$, $R_T = 0.34R_I$, $\mu_I = \mu_T = \mu_R = \frac{1}{3.3}$, $f_r = 0.002$, $\kappa = 0.0066$.

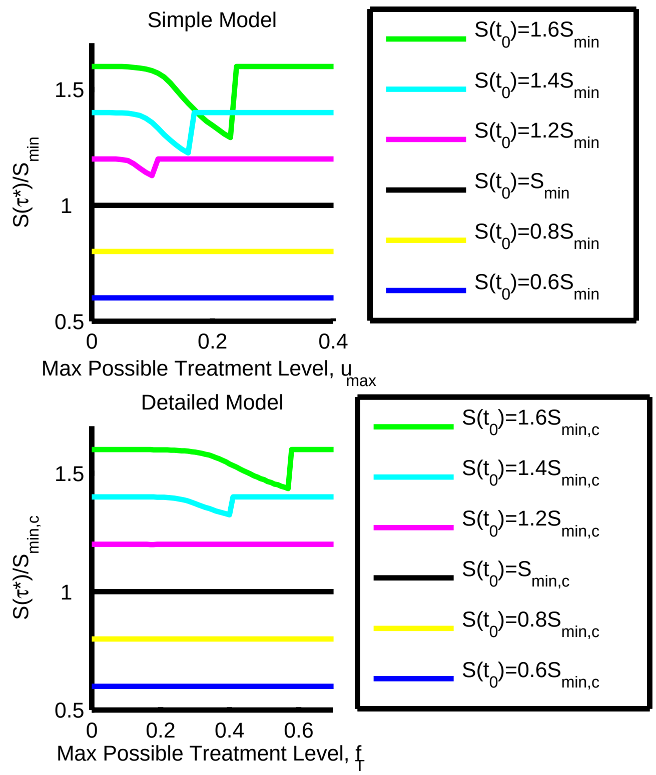

Figure 4.7: **Number of susceptibles at the optimal treatment start time ($S(\tau^*)$) versus treatment level.** Each curve corresponds to a different initial number of susceptibles ($S(t_0)$) **Panel A:** Figure was generated using simple model. $S_{\min}$ is the number of susceptibles at the optimal treatment start time when the treatment level is unbounded. If $S(t_0) \leq S_{\min}$ then it is always optimal to start treatment immediately. For example, the black curve corresponds to the case when $S(t_0) = S_{\min}$ and shows that for all values of $u_{\max}$ it is optimal to start treatment when $S(\tau^*) = S_{\min}$ (i.e., immediately). **Panel B:** Figure was generated using the more detailed model. $S_{\min,c}$ is the number of susceptibles at the optimal treatment start time assuming that once treatment begins all new infections are treated ($f_T = 1$). If $S(t_0) \leq S_{\min,c}$ then it is always optimal to start treatment immediately. Parameters are $R_R = 0.9R_I$, $R_T = 0.34R_I$, $\mu_I = \mu_T = \mu_R = \frac{1}{3.3}$, $f_r = 0.002$, $\kappa = 0.0066$.

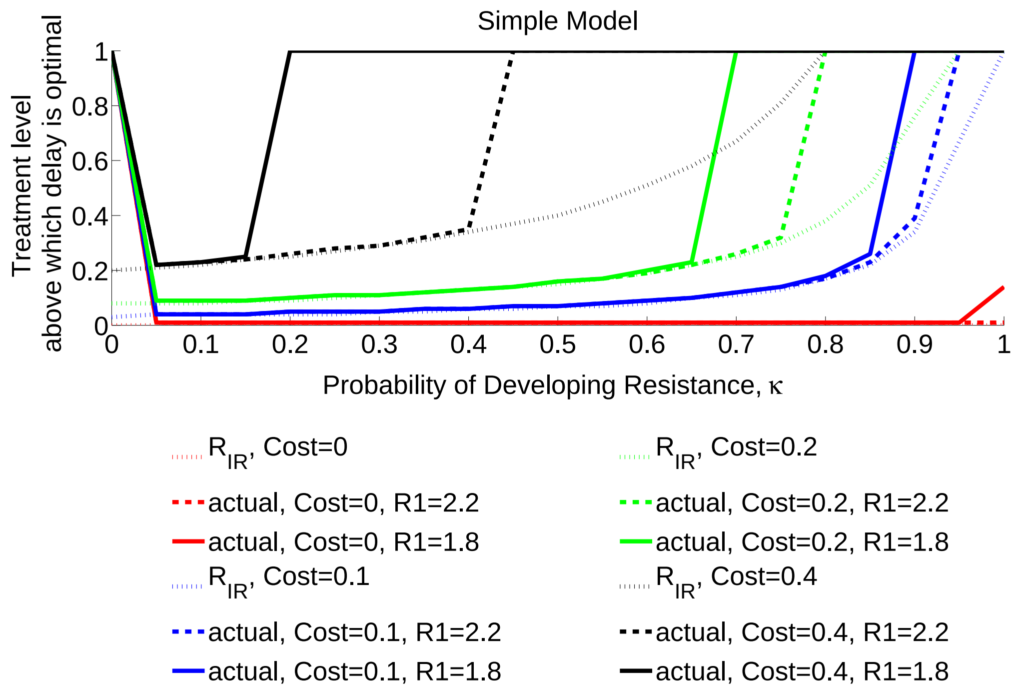

Figure 4.8: **Comparison of the theoretical result that teatment should begin immediately if $R_{IR} > R_R$ to actual simulation results.** Figure produced using simple model. R_{IR} is the expected number of secondary sensitive and resistant infections generated by an individual initially infected with the sensitive strain under constant maximal treatment. R_I, R_R are the basic reproduction number of the sensitive and resistant strain, respectively. Each color corresponds to a different fitness cost $(1 - \frac{R_R}{R_I})$. There are three curves for each color. Each dotted curve indicates the values of κ and $u_{\max}$ such that $R_{IR} = R_R$. By the theoretical result, for any point below the dotted curve it is optimal to start treatment immediately. Since our simulations adjust R_I by adjusting β_I, changing R_I will not change the position of the dotted curves. Each dashed curve corresponds to $R_I = 2.2$. Each dashed curve separates the $(\kappa, u_{\max})$ plane into two regions. Above the dashed curve some amount of delay is optimal. Below the dashed curve it is optimal to start treatment immediately. The solid curves are similar to the dashed curves except that they correspond to $R_I = 1.8$. Since for each color, the dashed and solid curves always lie above the dotted curve the theoretical results agree with the actual results. Parameters are $\mu_I = \mu_R = \frac{1}{3.3}$.

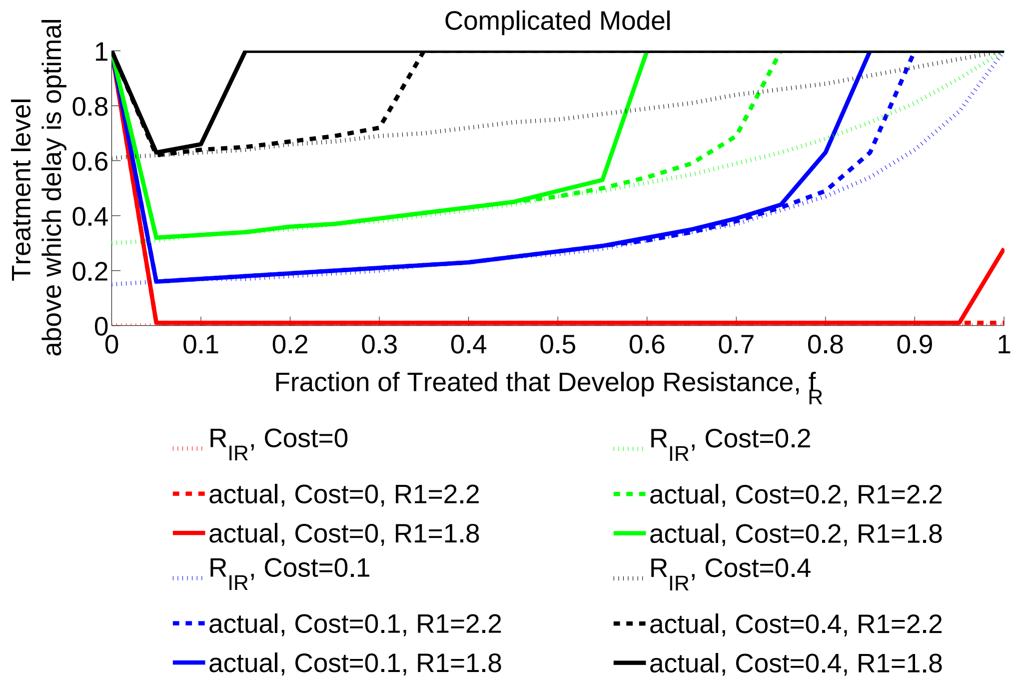

Figure 4.9: **Comparison of the theoretical result that teatment should begin immediately if $R_{IR} > R_R$ to actual simulation results.** Figure was produced using the detailed model and is analogous to Figure 4.8. R_{IR} is the expected number of secondary sensitive and resistant infections generated by an individual initially infected with the sensitive strain under constant maximal treatment. R_I, R_R are the basic reproduction number of the sensitive and resistant strain, respectively. Each color corresponds to a different fitness cost $(1 - \frac{R_R}{R_I})$. There are three curves for each color. Each dotted curve indicates the values of f_R and f_T such that $R_{IR} = R_R$. By the theoretical result, for any point below the dotted curve it is optimal to start treatment immediately. Since our simulations adjust R_I by adjusting β_I, changing R_I will not change the position of the dotted curves. Each dashed curve corresponds to $R_I = 2.2$. Each dashed curve separates the (f_R, f_T) plane into two regions. Above the dashed curve some amount of delay is optimal. Below the dashed curve it is optimal to start treatment immediately. The solid curves are similar to the dashed curves except that they correspond to $R_I = 1.8$. Since for each color, the dashed and solid curves always lie above the dotted curve the theoretical results agree with the actual results. Note that although the theoretical results were derived for the simple model they appear to also hold for the detailed model. Parameters are $\frac{\beta_T}{\mu_T} = 0.34R_I$, $\mu_I = \mu_T = \mu_R = \frac{1}{3.3}$.

4.6.6 Details of Proofs

When the Optimal Treatment Strategy is Singular

If the optimal control is singular on an interval $[t_1, t_2]$ then $\mathrm{SF}(t) = 0$ and $\dot{\mathrm{SF}}(t) = 0$ for $t \in [t_1, t_2]$. But this can only occur if

$$\dot{\lambda}_R = \left((S - S_R)(\kappa - 1) + \kappa(S_R - S_I)\right) \lambda_R \frac{\beta_I \beta_R I}{(\beta_I I + \beta_R R)^2 S} = 0$$

for all $t \in J$. But since S is a decreasing function of time this can only occur if either $\lambda_R = 0$ or $\kappa = 1$.

If $\lambda_R = 0$ then $\mathrm{SF} = 0$ implies that $\lambda_I = 0$ and therefore $t_f \in J$. But then, since $H = 0$ this implies that $\vec{\lambda}(t_f) = 0$ which is not possible.

If $\kappa = 1$ then in order for $\dot{\lambda}_R = 0$ we must have that $S_R = S_I$. Therefore, if the optimal control is singular it must be that $\kappa = 1$ and $S_R = S_I$.

We now solve for the singular control. Integrating the equations for model (4.11) from $t = t_0$ to $t = t_f$ and rearranging gives,

$$\begin{aligned} &S_I \ln(S(t_f)) - S(t_f) = \\ &\quad S_I \ln(S_0) - S_0 + I(t_f) + R(t_f) - I_0 - R_0 + (S_R - S_I)\, \beta_R \int_{t_0}^{t_f} R \mathrm{d}t + (1 - \kappa) \int_{t_0}^{t_f} uI \mathrm{d}t. \end{aligned} \tag{4.34}$$

Letting $\zeta = I_0 + R_0 - I(t_f) - R(t_f) > 0$, Equation (4.34) becomes

$$S_I \ln(S(t_f)) - S(t_f) = S_I \ln(S_0) - S_0 - \zeta + (S_R - S_I)\, \beta_R \int_{t_0}^{t_f} R \mathrm{d}t + (1 - \kappa) \int_{t_0}^{t_f} uI \mathrm{d}t, \tag{4.35}$$

Therefore, if $\kappa = 1$ and $(S_R - S_I) = 0$ then Equation (4.35) becomes

$$S_I \ln(S(t_f)) - S(t_f) = S_I \ln(S_0) - S_0 - \zeta, \tag{4.36}$$

This implies that any choice of u will result in an $S(t_f)$ that is a root of $f(S) = S_I \ln(S) - S$. Furthermore, since f is concave down, $S(t_f) < S_0$ and $\zeta > 0$, Equation (4.36) implies that $S(t_f)$ will always be the smaller root of f. That is, if $S_I = S_R$ and $\kappa = 1$ then any control is optimal and $S(t_f) < S_I$.

Monotonic Adjoint Functions

Claim 4.1. *The switch function* SF *is monotonic.*

Proof. First consider the case when u^* is singular. This occurs if $\kappa = 1$ and $S_I{=}S_R$. In this case, $\mathrm{SF}(t) = 0$ for all t and so the switch function is monotonic.

Suppose alternatively, that u^* is not singular. Suppose there exists a time $\hat{t}$ such that $\mathrm{SF}(\hat{t}) = 0$ and $\mathrm{sign}(\mathrm{SF})(\hat{t}^-) = \mathrm{sign}(\mathrm{SF})(\hat{t}^+)$. Then u^* can be chosen to be constant on some interval J containing $\hat{t}$ and therefore SF is analytic on J. Now, since SF is analytic on any interval where u^* is constant this means we can take higher order derivatives of SF at any singular point of SF.

Since $\dot{\mathrm{SF}} = \dot{\lambda}_R(\kappa + \frac{R}{I})$ we only need to show that λ_R is monotonic. Suppose $\dot{\lambda}_R = 0$ on an interval $[t_a, t_b]$. Then, by Equation (4.28), we have that (1) $\lambda_I(t) = \lambda_R(t)$ and (2) $\lambda_I(S_R - S_I - (1-\kappa)\frac{u}{\beta_I})|_t = 0$ for all $t \in [t_a, t_b]$. But this implies that $\dot{\lambda}_R(t) = 0$ for all $t > t_a$. Therefore if $\dot{\lambda}_R = 0$ on an interval, that interval includes t_f. This means that if λ_R is not monotonic then there exist times $t_i < t_f$ such that $\dot{\lambda}_R(t_i) = 0$ and $\mathrm{sign}(\dot{\lambda}_R(t_i^-)) = -\mathrm{sign}(\dot{\lambda}_R(t_i^+))$. Let t_1 be the largest such time.

Differentiating Equation (4.28) gives,

$$\ddot{\lambda}_R = \\ \left(\dot{S}(\lambda_I - \lambda_R) + (S - S_R)(\dot{\lambda}_I - \dot{\lambda}_R) + \dot{\lambda}_I(S_R - S_I) + (\dot{\lambda}_R\kappa - \dot{\lambda}_I)\frac{u}{\beta_I}\right)\frac{\beta_I\beta_R I}{(\beta_I I + \beta_R R)^2 S} \\ + \left((S - S_R)(\lambda_I - \lambda_R) + \lambda_I(S_R - S_I) + (\lambda_R\kappa - \lambda_I)\frac{u}{\beta_I}\right)\frac{\mathrm{d}}{\mathrm{d}t}\frac{\beta_I\beta_R I}{(\beta_I I + \beta_R R)^2 S} \tag{4.37}$$

Therefore,

$$\ddot{\lambda}_R(t_1) = \left(\dot{S}(t_1)(\lambda_I(t_1) - \lambda_R(t_1))\right)\frac{\beta_I\beta_R I(t_1)}{(\beta_I I(t_1) + \beta_R R(t_1))^2 S(t_1)}.$$

Therefore the concavity of λ_R at t_1 is determined by the sign of $(\lambda_I - \lambda_R)\big|_{t=t_1}$. That is, if $(\lambda_I(t_1) - \lambda_R(t_1)) < 0$ then λ_R (and hence SF) has a local minimum at t_1 and if $(\lambda_I(t_1) - \lambda_R(t_1)) > 0$ then λ_R (and SF) has a local maximum at t_1. That is,

$$\begin{aligned} \ddot{\lambda}_R(t_1) > 0 \quad &\text{if} \quad \lambda_I(t_1) < \lambda_R(t_1), \\ \ddot{\lambda}_R(t_1) = 0 \quad &\text{if} \quad \lambda_I(t_1) = \lambda_R(t_1), \\ \ddot{\lambda}_R(t_1) < 0 \quad &\text{if} \quad \lambda_I(t_1) > \lambda_R(t_1). \end{aligned} \tag{4.38}$$

Therefore, if $\lambda_I(t_1) < \lambda_R(t_1)$, then for all $t_1 < t < t_f$, $\lambda_R(t) > \lambda_R(t_1)$ and (since $\text{sign}(\dot{\lambda}_I) = -\text{sign}(\dot{\lambda}_R)$) $\lambda_I(t) < \lambda_I(t_1)$. That is, $\lambda_R(t_f) > \lambda_R(t_1) > \lambda_I(t_1) > \lambda_I(t_f)$. This contradicts the endpoint condition $\lambda_I(t_f) = \lambda_R(t_f)$ and so is not possible.

Similarly, if $\lambda_I(t_1) > \lambda_R(t_1)$, then for all $t_1 < t < t_f$, $\lambda_R(t) < \lambda_R(t_1)$ and $\lambda_I(t) > \lambda_I(t_1)$. That is, $\lambda_I(t_f) > \lambda_I(t_1) > \lambda_R(t_1) > \lambda_R(t_f)$. This contradicts the endpoint condition $\lambda_I(t_f) = \lambda_R(t_f)$ and so is not possible.

Lastly, consider the case when $\lambda_I(t_1) = \lambda_R(t_1)$ (i.e., when $\ddot{\lambda}_R(t_1) = 0$). From Equation (4.28) if $\lambda_I(t_1) = \lambda_R(t_1)$ then

$$\dot{\lambda}_R = \lambda_I(S_R - S_I - (1-\kappa)\frac{u}{\beta_I})\frac{\beta_I \beta_R I}{(\beta_I I + \beta_R R)^2 S}.$$

Therefore, if $\dot{\lambda}_R(t_1) = 0$ then either $\lambda_I(t_1) = 0$ or $S_R - S_I - (1-\kappa)\frac{u}{\beta_I} = 0$. But $\lambda_I = 0$ is not possible since it implies that $\vec{\lambda}(t_1) = 0$. Therefore, it must be that $S_R - S_I - (1-\kappa)\frac{u}{\beta_I} = 0$. But this implies that $\dot{\lambda}_I(t) = 0$ for all $t > t_1$ which, by the definition of t_1, is not possible. Therefore SF is monotonic.

$\square$

Table 4.1: Table of Symbols

Table of Symbols	
Symbol	Brief Definition
$S(t)$	number of susceptibles at time t;
$I(t)$	number of individuals infected with the treatment sensitive strain at time t;
$T(t)$	number of treated individuals at time t;
$R(t)$	number of individuals infected with the treatment resistant strain at time t;
β_I	contact parameter for the sensitive strain;
β_T	contact parameter for the treated strain ($\beta_T = 0$ for the simple model);
β_R	contact parameter for the resistant strain;
μ_I	per capita rate of death or recovery with immunity for the sensitive infected class;
μ_T	per capita rate of death or recovery with immunity for the treated class;
μ_R	per capita rate of death or recovery with immunity for the resistant infected class;
$u(t)$	per capita rate of treatment for the sensitive infected class at time t;
ν	per capita rate of developing resistance for the treated class (applies to model in Figure 4.1, Panel A);
κ	probability that treated individual develops resistance (applies to simple model);
t_0	the outbreak start time;
t_f	the outbreak end time;
$u_{\max}$	upper bound for per capita treatment rate u;
R_I	the basic reproduction number for the sensitive strain;
R_R	the basic reproduction number for the resistant strain;
R_{IR}	the expected number of secondary infections (both sensitive and resistant) caused by an individual initially infected with the sensitive strain, in a wholly susceptible population that is receiving maximum treatment (i.e., $u \equiv u_{\max}$);
S_I	the number of susceptibles above which a sensitive outbreak can occur;
S_R	the number of susceptibles above which a resistant outbreak can occur;
χ	the probability that an infected individual receives treatment (assuming $u \equiv u_{\max}$);
τ	an arbitrary treatment start time;
τ^*	the optimal treatment start time;
f_T	the fraction of infected individuals that receive treatment (applies to model in Figure 4.1, Panel A);
$S_{\min}$	the number of susceptibles at the optimal treatment start time assuming that an outbreak occurs and $u_{\max}$ is unbounded (applies to simple model,see Equation (4.1));
$S_{\min,\mathrm{c}}$	the number of susceptibles at the optimal treatment start time assuming that an outbreak occurs and $f_T = 1$ (applies to detailed model);
A	the total attack ratio.

Chapter 5

Resistance and the use of Multiple Drugs

5.1 Introduction

An important line of defense against drug resistance is the use of multiple drugs to treat a single disease. The general hypothesis is that by using multiple drugs the emergence and spread of resistance will be slowed. Even among those who agree with this hypothesis, there is disagreement as to how these drugs should be administered to the population. Two strategies that have garnered a lot of attention and study are the "cycling strategy" and the "mixed strategy". If a "cycling strategy" is used then at any time only one type of drug will be administered to the population and which drug is being administered will cycle in a periodic fashion. If a "mixed" strategy is used then at any particular time a certain fixed proportion of the infected population will be treated with each type of drug. Numerous variations on these two basic drug schedules have been considered. For example different types of alternating strategies (essentially aperiodic cycling strategies) and non-constant mixed strategies (where the proportion of the infected population receiving each type of drug is not fixed in time) but the basic "cycling" and "mixed" strategies have received the most attention. This is most likely due to the fact that these strategies are easily implemented and also lend themselves to analysis via numerical simulation. In this chapter we consider the optimal use of two drugs to minimize the infectious burden of an endemic disease. As a motivating example we will review the influential paper by Bonhoeffer et al. [14] and then consider two simplified models.

5.2 A Motivating Example

In [14], the following model of the spread of a hospital infection was considered:

$$\begin{aligned}
\dot{S} &= M - \mu_S S - \beta(I_s + I_1 + I_2)S + r_1 I_1 + r_2 I_2 + (1-\kappa)(u_m I_s + u I_1 + (u_m - u)I_2), \\
\dot{I}_s &= (\beta S - \mu_I - u_m)I_s, \\
\dot{I}_1 &= (\beta S - \mu_I - r_1 - u)I_1 + \kappa(u_m - u)I_s, \\
\dot{I}_2 &= (\beta S - \mu_I - r_2 - u_m + u)I_2 + \kappa u I_s.
\end{aligned} \tag{5.1}$$

In model (5.1) susceptible individuals enter the hospital at a constant rate M and are discharged at the per capita rate μ_S. The rate at which susceptibles become infected is governed by a mass action law with contact parameter β and these infections can be treated with two different drugs (Drug 1 or Drug 2). There are three different strains that can infect a susceptible: a strain I_s that is susceptible to both drugs, a strain I_1 that is susceptible to Drug 1 and resistant to Drug 2 and finally a strain I_2 that is susceptible to Drug 2 but resistant to Drug 1. If an individual infected with the I_s strain is treated with Drug i they will either recover and re-enter the susceptible compartment (with probability $1-\kappa$) or will develop resistance to Drug i (with probability κ). Individuals infected with the I_i strain can clear the infection at a per capita rate of r_i and re-enter the susceptible compartment. If an individual infected with the I_i strain is treated with Drug i then they will either recover and re-enter the susceptible compartment (with probability $1-\kappa$) or they will develop resistance to Drug i (with probability κ). It is assumed that a strain that is resistant to a single drug suffers a fitness cost which makes it possible for the infected individual to clear the infection without treatment (hence the per capita recovery rates r_1 and r_2). Furthermore it is assumed that any strain that is resistant to both drugs suffers a sufficiently high fitness cost that individuals infected with this strain are not infectious (in [14] a second model that includes infectious individuals that are resistant to both drugs is also considered however we will not discuss this model here). Infected individuals die at a per capita rate μ_I. A few important things to note about this model is that both drugs are equally effective on strain I_s and that the behaviour of the two strains I_1 and I_2 differ only with respect to their recovery rates r_1 and r_2. Finally, for ease of comparison with other models we have assumed that at any instant in time the maximum number of I_s infections are always treated (i.e., the number of I_s infections being treated at any time is $u_m I_s$). Although this assumption is not explicitly made in [14], this assumption does not change any of the analysis in [14].

Using model (5.1) and under the assumption that $r_1 = r_2$ and $\kappa = 0$, Bonhoeffer et al. used numerical simulations to compare (i) a “50-50 mixed strategy” where 50 percent of the infected population received Drug 1 and 50 percent of the infected population received Drug 2 to (ii) a cycling strategy with a cycling period of 5 time units and found that the net gain of susceptibles over a fixed time period was larger for the mixed strategy than for the cycling strategy. Although the simulations presented in [14] are restricted to the case when $r_1 = r_2$, they also discuss how the "50-50 mixed" strategy could be modified to account for asymmetries in the resistant strains. Specifically, if $r_2 > r_1$ then they propose that the "50-50 mixed" strategy should be replaced with a mixed strategy where the fraction of infecteds treated with Drug 2 is $u = \frac{r_1 - r_2 + u_m}{2}$. We will refer to this treatment strategy as the 'adjusted mixed strategy'. Using the adjusted mixed strategy ensures that the time required for either resistant strain to increase from an initial frequency ρ_0 to a frequency ρ_e is the same. Two interesting questions that arise from the analysis in [14] are:

(i) If more general types of treatment strategies are considered is a mixing strategy still best?

(ii) How does de novo resistance affect the form of the best treatment strategy?

A preliminary investigation indicates that the answers to these questions may indeed result in some unexpected conclusions. In Figure 1 we compare the adjusted mixed strategy to all possible mixed strategies for different probabilities of de novo resistance for model (5.1). The top row of Figure 1 assumes that the the two resistant strains are identical ($r_1 = r_2$) and the bottom row assumes $r_1 \neq r_2$. It is interesting to note that the adjusted mixed strategy is the best mixed strategy for all values of κ when $r_1 = r_2$ but if the two resistant strains differ then the adjusted mixed strategy is the best mixed strategy only when $\kappa = 0$.

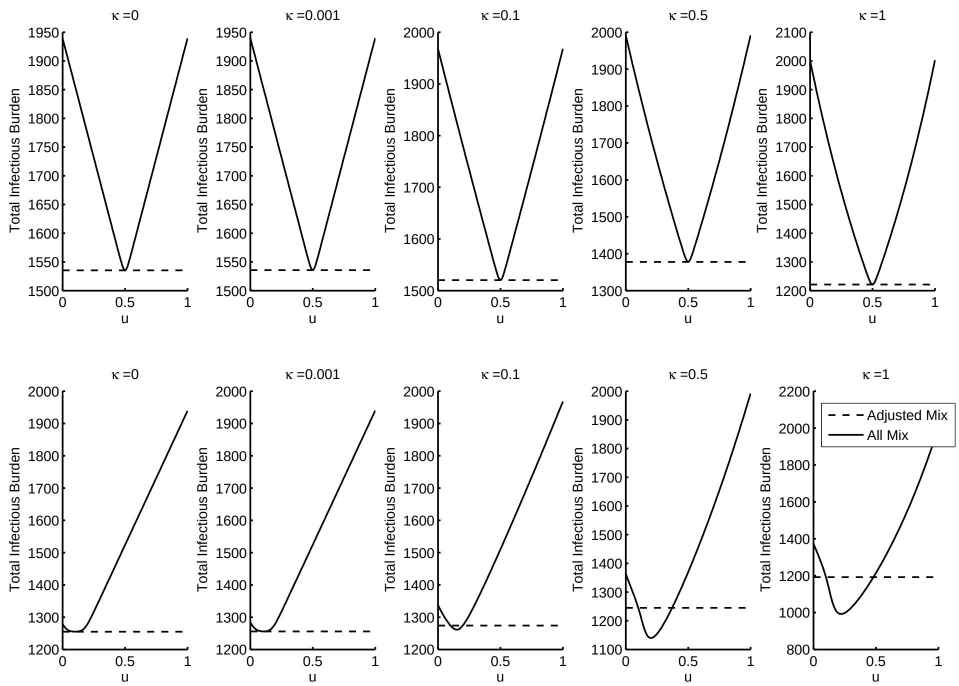

Figure 5.1: Comparison of the adjusted mixed strategy to all other mixed strategies. **Top row**: Both resistant strains are identical ($r_1 = r_2$). The dashed line lies below the solid curve therefore the adjusted mixed strategy ($u = \frac{u_m}{2}$) is the best constant mixing strategy for all values of κ. **Bottom row**: The resistant strains are not identical ($r_1 > r_2$). The dashed line lies below the solid curve only when $\kappa = 0$. Therefore the adjusted mixed strategy ($u = \frac{r_1 - r_2 + u_m}{2}$) is the best constant mixed strategy only when $\kappa = 0$. Furthermore, increasing κ decreases the performance of $u = \frac{r_1 - r_2 + u_m}{2} + \frac{u_m}{2}$. Parameter values are those used in [14] (i.e., $M = 100$, $\mu_s = 1$, $\mu_I = 1.5$, $r_1 = r_2 = 0.1$ and $u_m = 1$) except in the second row where $r_1 = 0.9$ and $r_2 = 0.1$.

5.3 Optimal Treatment When Treatment Does Not Cause Resistance: A Simple Model

As a first step towards answering the question of whether or not mixing is actually the best strategy we begin by considering a simple two strain model that does not include de novo resistance. Although the model we consider is simpler than the model discussed in [14] it still captures a number of the important aspects of a two strain endemic disease. Furthermore, the model is simple enough that it can be analytically solved using Pontryagin's Maximum Principle.

5.3.1 Model Description

We consider the following model,

$$\begin{aligned} \dot{I}_s &= (r_s - u_m)I_s + \phi, \\ \dot{I}_1 &= (r_1 - u)I_1 + \phi, \\ \dot{I}_2 &= (r_2 - u_m + u)I_2 + \phi, \end{aligned} \tag{5.2}$$

where r_s, r_1, $r_2 < 0$, ϕ, $u_m > 0$ are constants and $u : [t_0, t_f] \to [0, u_m]$. In model (5.2) there are three different strains: a strain I_s that is susceptible to both drugs, a strain I_1 that is susceptible to Drug 1 and resistant to Drug 2 and finally a strain I_2 that is susceptible to Drug 2 but resistant to Drug 1. The quantity $u(t)$ denotes the fraction of I_s infections that are being treated with Drug 1 at time t. Furthermore, we assume that both drugs are equally effective on I_s infections and that at any instant in time the maximum possible number of I_s infections are treated (i.e,. the number of I_s infections being treated with Drug 2 at time t is $u_m - u$). Under these assumptions the dynamics of the I_s strain are unaffected by changing the control strategy u and hence in the sequel we will only consider the final two equations of model (5.2). New I_1 infections enter the system at a constant rate ϕ and I_1 individuals either die or recover with immunity at a per capita rate r_1. I_1 individuals are successfully treated (and acquire immunity) at a time-varying per capita rate u. Similar dynamics hold for the I_2 strain. Notice that the rate at which individuals leave their infected class is modulated by the treatment strategy. For example, if the treatment effort is entirely targeted at strain 1 then $u = u_m$ and the per capita outflux of strain 1 and strain 2 is $r_1 - u_m$ and r_2, respectively. Alternatively, if the treatment effort is entirely targeted at strain 2 then $u = 0$ and the per capita outflux of strain 1 and strain 2 is r_1 and $r_2 - u_m$, respectively. Any intermediate value of u indicates a treatment strategy that targets both strains. Our ultimate goal is to describe the function u that minimizes the total infectious burden.

5.3.2 Optimal Treatment Strategy

Problem 5.1. *Fix $t_f > t_0$ and $u_m > 0$. Find the function $u : [t_0, t_f] \to [0, u_m]$ that minimizes the total infectious burden*

$$\int_{t_0}^{t_f} (I_1(t) + I_2(t))\mathrm{d}t$$

for model (5.2).

Theorem 5.1. *Let $\bar{u} = \frac{r_1 - r_2 + u_m}{2}$. Assume that $r_2 - u_m \leq r_1 - u_m < r_2 \leq r_1$. Then* *there are three possibilities:*

(i) If $I_2(t_0) = I_1(t_0)$ *then* $u^* \equiv \bar{u}$.

(ii) If $I_2(t_0) > I_1(t_0)$ *then*

$$u^*(t) = \begin{cases} 0 & \text{for } t \in [0, \tilde{t}] \\ \bar{u} & \text{for } t \in (\tilde{t}, t_f], \end{cases}$$

where $\tilde{t} = \min\{t_f, \min\{t \mid I_2(t) = I_1(t)\}\}$.

(iii) If $I_2(t_0) < I_1(t_0)$ *then*

$$u^*(t) = \begin{cases} u_m & \text{for } t \in [0, \tilde{t}] \\ \bar{u} & \text{for } t \in (\tilde{t}, t_f], \end{cases}$$

where $\tilde{t} = \min\{t_f, \min\{t \mid I_2(t) = I_1(t)\}\}$.

The assumption that $r_1 - u_m < r_2$ ensures that if strain 1 is targeted exclusively then the per capita rate of outflux of strain 1 is greater than that of strain 2. The assumption that $r_2 - u_m < r_1$ has an analogous interpretation. Theorem 5.1 states that the optimal treatment strategy targets the most prevalent strain until both strains are equally prevalent. Once the two strains are present in equal amounts the control switches to an intermediate treatment value to ensure they remain equally prevalent.

Proof. The proof of Theorem 5.1 uses Pontryagin's Maximum Principle (PMP). The existence of an optimal control can be shown using a straightforward application of Fillipov's theorem. The relations provided by PMP are as follows:

The Hamiltonian is

$$H = \lambda_0(I_1 + I_2) + \lambda_1(r_1 - u)I_1 + \lambda_1\phi + \lambda_2(r_2 - u_m + u)I_2 + \lambda_2\phi.$$

The adjoint variables are described by the following equations

$$\begin{aligned} \dot{\lambda}_1 &= -\lambda_0 - \lambda_1(r_1 - u), \\ \dot{\lambda}_2 &= -\lambda_0 - \lambda_2(r_2 - u_m + u), \end{aligned} \tag{5.3}$$

with transversality conditions $(\lambda_0, \lambda_1(t_f), \lambda_2(t_f)) = (-1, 0, 0)$. By Equations (5.3) and the transversality condition $(\lambda_0, \lambda_1(t_f), \lambda_2(t_f)) = (-1, 0, 0)$ we see that $\lambda_1(t)$, $\lambda_2(t) < 0$ for all $t \in [t_0, t_f)$. The switch function is $\mathrm{SF} = \lambda_2 I_2 - \lambda_1 I_1$.

Claim 5.1. *Let u^* denote the optimal treatment strategy and I_1^* and I_2^* the corresponding optimal trajectories. If the optimal treatment strategy is singular on an interval $J = [t_1, t_2]$ then $u^*(t) = \bar{u}$ for all $t \in J$. Furthermore, $I_1^*(t) = I_2^*(t)$ for all $t \in J$.*

Proof. In order for u^* to be singular on an interval J we must have that $\mathrm{SF} = 0$ and $\dot{\mathrm{SF}} = \lambda_0(I_1 - I_2) - \phi(\lambda_1 - \lambda_2) = 0$. Substituting $\mathrm{SF} = 0$ into $\dot{\mathrm{SF}} = 0$ gives

$$0 = (\lambda_0 + \phi\frac{\lambda_2}{I_1})(I_1 - I_2).$$

Therefore, if the control is singular on an interval J then either $I_1^*(t) = I_2^*(t)$ or $\lambda_0 + \phi\frac{\lambda_2(t)}{I_1^*(t)} = 0$ for all $t \in J$.

Suppose $I_1^*(t) = I_2^*(t)$ for all $t \in J$. Then $\dot{I}_1^* = \dot{I}_2^*$ on J which implies that

$u^* = \frac{r_1 - r_2 + u_m}{2} = \bar{u}$ on J. Alternatively, suppose $\lambda_0 + \phi \frac{\lambda_2}{I_1^*} = 0$ on J. This implies that

$$\lambda_0 \dot{I}_1 = -\phi \dot{\lambda}_2. \tag{5.4}$$

Substituting the expression for $\dot{I}_1$ and $\dot{\lambda}_2$ into Equation (5.4) gives,

$$\lambda_0 (r_1 - u) I_1 + \lambda_0 \phi = \phi \lambda_0 + \phi \lambda_2 (r_2 - u_m + u). \tag{5.5}$$

Substituting $\lambda_0 + \phi \frac{\lambda_2}{I_1} = 0$ into Equation (5.5) gives $0 = \lambda_0 I_1 (r_2 + r_1 - u_m)$ which is not possible since $\lambda_0 = -1$. Therefore, if there is a singular control on an interval J then on this interval $I_1^* = I_2^*$, $\lambda_1 = \lambda_2$ and $u^* = \frac{r_1 - r_2 + u_m}{2} = \bar{u}$. $\square$

By Claim 5.1 we know that the interval $[t_0, t_f]$ can be partitioned into maximal disjoint subintervals J_i such that on the i^{th} subinterval $u^* = c_i$ where $c_i \in \{0, u_{\max}, \bar{u}\}$. In the sequel we will denote the final interval by $J_f = (t_1, t_f]$. We will use the following claim:

Claim 5.2. *Suppose there exist times t_1 and t_2 such that (i) $t_2 > t_1 \geq t_0$, (ii) $u(t) = 0$ for all $t \in (t_1, t_2]$ and (iii) $I_2(t_2) \geq I_1(t_2)$. Then $I_2(t) > I_1(t)$ for all $t \in [t_1, t_2)$. Alternatively, if (i) $t_2 > t_1 \geq t_0$, (ii) $u(t) = u_m$ for all $t \in (t_1, t_2]$ and (iii) $I_2(t_2) \leq I_1(t_2)$ then $I_2(t) < I_1(t)$ for all $t \in [t_1, t_2)$.*

Proof. Proof by contradiction. Assume (i) $t_2 > t_1 \geq t_0$, (ii) $u(t) = 0$ for all $t \in (t_1, t_2]$ and (iii) $I_2(t_2) \geq I_1(t_2)$ and suppose there exist times $\bar{t}_i \in [t_1, t_2)$ such that $I_2(\bar{t}_i) = I_1(\bar{t}_i)$. Let $\bar{t}_n$ be the largest such time. Then $\dot{I}_1(\bar{t}_n) > \dot{I}_2(\bar{t}_n)$ which implies that $I_1(t) > I_2(t)$ for all $t \in (\bar{t}_n, t_2]$. This contradicts the assumption that $I_2(t_2) > I_1(t_2)$. An analogous argument holds for the case when (i) $t_2 > t_1 \geq t_0$, (ii) $u(t) = u_m$ for all

$t \in [t_1, t_2]$ and (iii) $I_2(t_2) < I_1(t_2)$. $\square$

The remainder of the proof of Theorem 5.1 can be separated into three different cases:

Case 1: Suppose that $I_2^*(t_f) > I_1^*(t_f)$. From the transversality conditions we know

$$\dot{\mathrm{SF}}(t_f) = \lambda_0(I_1^*(t_f) - I_2^*(t_f)) = (I_2^*(t_f) - I_1^*(t_f)) > 0$$

and so $u^*(t) = 0$ for all $t \in J_f = (t_1, t_f]$. Suppose $t_1 > t_0$ then $\mathrm{SF}(t_1) = 0$ and

$$\dot{\mathrm{SF}}(t_1) = (I_2(t_1) - I_1(t_1)) - \phi(\lambda_1(t_1) - \lambda_2(t_1)) \leq 0. \tag{5.6}$$

By Claim 5.2 we know that $I_2^*(t_1) > I_1^*(t_1)$ and therefore Equation (5.6) implies that $\lambda_1(t_1) \geq \lambda_2(t_1)$. But this is not possible since $\mathrm{SF}(t_1) = \lambda_2(t_1)I_2(t_1) - \lambda_1(t_1)I_1(t_1) = 0$. Therefore, if $I_2^*(t_f) > I_1^*(t_f)$ then $I_2^*(t_0) > I_1^*(t_0)$ and the optimal control is $u^* \equiv 0$.

Case 2: Suppose that $I_1^*(t_f) > I_2^*(t_f)$. From the transversality conditions we know

$$\dot{\mathrm{SF}}(t_f) = \lambda_0(I_1^*(t_f) - I_2^*(t_f)) = (I_2^*(t_f) - I_1^*(t_f)) < 0$$

and so $u^*(t) = u_m$ for all $t \in J_f = (t_1, t_f]$. Suppose $t_1 > t_0$ then $\mathrm{SF}(t_1) = 0$ and $\dot{\mathrm{SF}}(t_1) = (I_2(t_1) - I_1(t_1)) - \phi(\lambda_1(t_1) - \lambda_2(t_1)) \geq 0$. By Claim 5.2 we know that $I_1^*(t_1) > I_2^*(t_1)$ and therefore $\dot{\mathrm{SF}}(t_1) \geq 0$ implies that $\lambda_1(t_1) \leq \lambda_2(t_1)$. But this is not possible since $\mathrm{SF}(t_1) = \lambda_2(t_1)I_2(t_1) - \lambda_1(t_1)I_1(t_1) = 0$. Therefore, if $I_1^*(t_f) > I_2^*(t_f)$ then $I_1^*(t_0) > I_2^*(t_0)$ and the optimal control is $u^* \equiv u_m$.

Case 3: Suppose that $I_2^*(t_f) = I_1^*(t_f)$. Notice that in this case we have $\mathrm{SF}(t_f) = 0$ and $\dot{\mathrm{SF}}(t_f) = 0$. There are three possibilities:

(i) The first possibility is that

$$\begin{aligned}\ddot{\mathrm{SF}}(t_f) &= [(r_2 - u_m + u)I_2 - (r_1 - u)I_1 + \phi((r_1 - u)\lambda_1 - (r_2 - u_m + u)\lambda_2)]_{(t=t_f)}, \\ &= I_1(t_f)(r_2 - r_1 - u_m + 2u) > 0.\end{aligned}$$

In this case we can make the following observations:

(a) Since $\ddot{\mathrm{SF}}(t_f) \neq 0$, u^* is not singular on J_f and therefore $u^* \in \{0, u_m\}$ on J_f.

(b) $\ddot{\mathrm{SF}}(t_f) > 0$ implies that $u > \bar{u}$ and therefore $u^* = u_m$ on J_f.

Suppose $t_1 > t_0$ then

$$\dot{\mathrm{SF}}(t_1) = (I_2(t_1) - I_1(t_1)) - \phi(\lambda_1(t_1) - \lambda_2(t_1)) \geq 0.$$

Furthermore, by Claim 5.2, $I_1^*(t_1) > I_2^*(t_1)$ and therefore $\dot{\mathrm{SF}}(t_1) \geq 0$ is only possible if $\lambda_1(t_1) < \lambda_2(t_1) < 0$. But this is not possible since (i) $\mathrm{SF} = \lambda_2 I_2 - \lambda_1 I_1 = 0$ and (ii) $I_1 > I_2$ imply $|\lambda_2(t_1)| > |\lambda_1(t_1)|$. Therefore if $I_2^*(t_f) = I_1^*(t_f)$ and $\ddot{\mathrm{SF}}(t_f) > 0$ then $u^* \equiv u_m$ and $I_1(t_0) > I_2(t_0)$.

(ii) The second possibility is that

$$\begin{aligned}\ddot{\mathrm{SF}}(t_f) &= [(r_2 - u_m + u)I_2 - (r_1 - u)I_1 + \phi((r_1 - u)\lambda_1 - (r_2 - u_m + u)\lambda_2)]_{(t=t_f)}, \\ &= I_1(t_f)(r_2 - r_1 - u_m + 2u) < 0.\end{aligned}$$

In this case $r_2 - r_1 - u_m + 2u < 0$ and so $u^* = 0$ on J_f. Suppose $t_1 > t_0$ then

$$\dot{\mathrm{SF}}(t_1) = (I_2(t_1) - I_1(t_1)) - \phi(\lambda_1(t_1) - \lambda_2(t_1)) \leq 0.$$

Furthermore, by Claim 5.2, $I_1^*(t_1) < I_2^*(t_1)$ and therefore $\dot{\mathrm{SF}}(t_1) \leq 0$ is only possible if $\lambda_2(t_1) < \lambda_1(t_1) < 0$. But this is not possible since (i) $\mathrm{SF} = \lambda_2 I_2 - \lambda_1 I_1 = 0$ and (ii) $I_1 < I_2$ imply $|\lambda_1(t_1)| > |\lambda_2(t_1)|$. Therefore if $I_2^*(t_f) = I_1^*(t_f)$ and $\ddot{\mathrm{SF}}(t_f) < 0$ then $u^* \equiv 0$ and $I_1(t_0) < I_2(t_0)$.

(iii) The third possibility is that

$$\begin{aligned}
\ddot{\mathrm{SF}}(t_f) &= [(r_2 - u_m + u)I_2 - (r_1 - u)I_1 + \phi((r_1 - u)\lambda_1 - (r_2 - u_m + u)\lambda_2)]_{(t=t_f)}, \\
&= I_1(t_f)(r_2 - r_1 - u_m + 2u) = 0.
\end{aligned}$$

This is only possible if $u = \bar{u}$ on $J_f = (t_1, t_f]$. If $t_1 = t_0$ then $u^* = \bar{u}$ and $I_1(t_0) = I_2(t_0)$. Alternatively, if $t_1 > t_0$ then $\mathrm{SF}(t_1) = 0$, $I_2^*(t_1) = I_1^*(t_1)$ and $\lambda_2^*(t_1) = \lambda_1^*(t_1)$. Therefore $\dot{\mathrm{SF}}(t_1) = 0$ and

$$\ddot{\mathrm{SF}}(t_1) = (I_1(t_1) - \lambda_1(t_1)\phi)(r_2 - r_1 - u_m + 2u).$$

If $\ddot{\mathrm{SF}}(t_1) > 0$ then the proof of "the first possibility" applies (replace t_1 with t_2 and t_f with t_1) and therefore $u^*(t) = u_m$ for all $t \in [t_0, t_1]$ and $I_1(t_0) > I_2(t_0)$. Alternatively, if $\ddot{\mathrm{SF}}(t_1) < 0$ then the proof of "the second possibility" applies (replace t_1 with t_2 and t_f with t_1) and therefore $u^*(t) = 0$ for all $t \in [t_0, t_1]$ and $I_2(t_0) > I_1(t_0)$.

$\square$

5.4 Optimal Treatment When Treatment Can Cause Resistance: A Simple Model

5.4.1 Model Description

Previously, we showed that the treatment strategy that minimizes the total infectious burden for

$$\begin{aligned} \dot{I}_1 &= (r_1 - u)I_1 + \phi, \\ \dot{I}_2 &= (r_2 - u_m + u)I_2 + \phi, \end{aligned} \tag{5.7}$$

is to target the most prevalent strain until both strains are equally prevalent and then use a singular control that ensures that the two strains remain equally prevalent. Let I_s denote a strain that is sensitive to both drugs. Then clearly, this result also holds for the following system (since the I_s dynamics are unaffected by changing the treatment strategy):

$$\begin{aligned} \dot{I}_s &= (r_s - u_m)I_s + \phi_s, \\ \dot{I}_1 &= (r_1 - u)I_1 + \phi, \\ \dot{I}_2 &= (r_2 - u_m + u)I_2 + \phi. \end{aligned} \tag{5.8}$$

The next step for our analysis is to consider the effect of de novo resistance. To begin, consider the following model:

$$\begin{aligned}
\dot{I}_s &= (r_s - u_m)I_s + \phi_s, \\
\dot{I}_1 &= (r_1 - u)I_1 + \phi + \kappa(u_m - u)I_s, \\
\dot{I}_2 &= (r_2 - u_m + u)I_2 + \phi + \kappa u I_s.
\end{aligned} \tag{5.9}$$

In model (5.9) I_s individuals that are treated with Drug 1 develop resistance to Drug 1 (and hence enter the I_2 compartment) with probability κ. Similarly, I_s individuals that are treated with Drug 2 enter the I_1 compartment with probability κ. Notice that I_s will eventually reach the equilibrium $I_s = \frac{-\phi_s}{r_s - u_m} = \alpha\phi$. Therefore, once I_s reaches equilibrium model (5.9) becomes

$$\begin{aligned}
\dot{I}_s &= 0, \\
\dot{I}_1 &= (r_1 - u)I_1 + \phi + \kappa(u_m - u)\alpha\phi, \\
\dot{I}_2 &= (r_2 - u_m + u)I_2 + \phi + \kappa u \alpha\phi.
\end{aligned} \tag{5.10}$$

Redefining κ to be $\kappa\alpha$ we see that finding the optimal treatment strategy for model (5.10) is equivalent to finding the optimal treatment strategy for

$$\begin{aligned}
\dot{I}_1 &= (r_1 - u)I_1 + (1 + \kappa(u_m - u))\phi, \\
\dot{I}_2 &= (r_2 - u_m + u)I_2 + (1 + \kappa u)\phi.
\end{aligned} \tag{5.11}$$

5.4.2 Optimal Treatment Strategy

Problem 5.2. *Fix $t_f > t_0$ and $u_m > 0$ and assume $0 > r_1 \geq r_2 \geq r_1 - u_m \geq r_2 - u_m$. Find the function $u : [t_0, t_f] \to [0, u_m]$ that minimizes the total infectious burden*

$$\int_{t_0}^{t_f} (I_1(t) + I_2(t))\mathrm{d}t$$

for model (5.11).

Although a complete analytic solution to Problem 5.2 remains elusive a number of important observations can be made by examining the solution to Problem 5.2 under the assumption that $r_1 = r_2$.

Theorem 5.2. *If $r_1 = r_2$ then the solution to Problem 5.2 has the same form as the solution to Problem 5.1. Specifically,*

(i) If $I_2(t_0) = I_1(t_0)$ then $u^* \equiv \bar{u}$.

(ii) If $I_2(t_0) > I_1(t_0)$ then

$$u^*(t) = \begin{cases} 0 & \text{for } t \in [0, \tilde{t}] \\ \bar{u} & \text{for } t \in (\tilde{t}, t_f], \end{cases}$$

where $\tilde{t} = \min\{t_f, \min\{t \mid I_2(t) = I_1(t)\}\}$.

(iii) If $I_2(t_0) < I_1(t_0)$ then

$$u^*(t) = \begin{cases} u_m & \text{for } t \in [0, \tilde{t}] \\ \bar{u} & \text{for } t \in (\tilde{t}, t_f], \end{cases}$$

where $\tilde{t} = \min\{t_f, \min\{t \mid I_2(t) = I_1(t)\}\}$.

Proof. The proof of Theorem 5.2 uses PMP. The existence of an optimal control can be shown using a straightforward application of Fillipov's theorem. The relations provided by PMP are as follows.

The Hamiltonian is

$$H = \lambda_0(I_1+I_2)+\lambda_1(r_1-u)I_1+\lambda_1\phi+\lambda_1\kappa\phi(u_m-u)+\lambda_2(r_2-u_m+u)I_2+\lambda_2\phi+\lambda_2\kappa u\phi.$$

The adjoint variables are defined by

$$\begin{aligned} \dot{\lambda}_1 &= -\lambda_0 - \lambda_1(r_1 - u), \\ \dot{\lambda}_2 &= -\lambda_0 - \lambda_2(r_2 - u_m + u), \end{aligned} \tag{5.12}$$

with transversality conditions $(\lambda_0, \lambda_1(t_f), \lambda_2(t_f)) = (-1, 0, 0)$. The switch function is

$$\mathrm{SF} = \lambda_2(I_2 + \kappa\phi) - \lambda_1(I_1 + \kappa\phi).$$

Furthermore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{SF} = I_2 - I_1 + \lambda_1\phi[\kappa(r_1 - u_m) - 1] - \lambda_2\phi[\kappa(r_2 - u_m) - 1].$$

From the transversality conditions we have that $\mathrm{SF}(t_f) = 0$ and $\dot{\mathrm{SF}}(t_f) = I_2(t_f) - I_1(t_f)$.

Claim 5.3. *Assume* $r_1 = r_2$. *If* u^* *is singular on an interval* J *then* $I_1(t) = I_2(t)$ *and* $u^*(t) = \bar{u} = \frac{u_m}{2}$ *for all* $t \in J$.

Proof. If u^* is singular on the interval J then $\mathrm{SF}(t) = 0$ and $\dot{\mathrm{SF}}(t) = 0$ for all $t \in J$. Substituting $\mathrm{SF}(t) = 0$ into $\dot{\mathrm{SF}}(t) = 0$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{SF} = (I_2 - I_1)(1 + \lambda_1 \phi \frac{[\kappa(r_1 - u_m) - 1]}{I_2 + \kappa\phi}).$$

Now since $\lambda_1 \leq 0$ (see Remark 5.1), we have that $1+\lambda_1\phi\frac{[\kappa(r_1-u_m)-1]}{I2+\kappa\phi} \geq 0$ and therefore $\dot{\mathrm{SF}} = 0$ on J if and only if $I_2 = I_1$ on J. Finally, $I_2(t) = I_1(t)$ for all $t \in J$ implies that $\dot{I}_2 = \dot{I}_1$ on J which implies that $u^* = \frac{u_m}{2}$.

Remark 5.1. *By Equation* (5.12) *for any* $\bar{t}$ *such that* $\lambda_1(\bar{t}) = 0$ *we have that* $\dot{\lambda}_1(\bar{t}) > 0$. *Therefore, since* $\lambda_1(t_f) = 0$, *it must be that* $\lambda_1 < 0$ *for all* $t \in [t_0, t_f)$.

□

By Claim 5.3 we know that the interval $[t_0, t_f]$ can be partitioned into disjoint subintervals J_i such that for each i, $u^* \equiv c_i$ on J_i and $c_i \in \{0, u_m, \frac{u_m}{2}\}$. We will use the following claim:

Claim 5.4. *Fix* $t_1 < t_2$. *If* $I_2(t_2) \geq I_1(t_2)$ *and* $u(t) = 0$ *for all* $t \in (t_1, t_2]$ *then* $I_2(t) > I_1(t)$ *for all* $t \in [t_1, t_2)$. *Similarly, if* $I_1(t_2) \geq I_2(t_2)$ *and* $u(t) = u_m$ *for all* $t \in (t_1, t_2]$ *then* $I_1(t) > I_2(t)$ *for all* $t \in [t_1, t_2]$.

Proof. Assume $I_2(t_2) \geq I_1(t_2)$ and $u(t) = 0$ for all $t \in (t_1, t_2]$. Suppose there exist times $\tau_i \in [t_1, t_2)$ such that $I_1(\tau_i) = I_2(\tau_i)$. Let τ_1 be the largest such time. Then $\dot{I}_1(\tau_1) > \dot{I}_2(\tau_1)$. But this implies that $I_1(t_2) > I_2(t_2)$. An analogous argument holds for the case when $I_1(t_2) > I_2(t_2)$ and $u(t) = u_m$ for all $t \in (t_1, t_2]$. □

The remainder of the proof of Theorem 5.2 can be separated into three cases:

Case 1: Suppose $I_2(t_f) > I_1(t_f)$. Then

$$\dot{\mathrm{SF}}(t_f) = (I_2(t_f) - I_1(t_f))(1 + \lambda_1(t_f)\phi\frac{\kappa(r_1 - u_m) - 1}{I_2(t_f) + \kappa\phi}) = (I_2(t_f) - I_1(t_f) > 0.$$

Therefore, $u = 0$ on $J_f = (t_1, t_f]$ and by Claim 5.4 we have $I_2(t_1) > I_1(t_1)$. Suppose $t_1 > t_0$, then we have that $\dot{\mathrm{SF}}(t_1) < 0$. But this is not possible since $\dot{\mathrm{SF}}(t_1) = (I_2(t_1) - I_1(t_1))(1 + \lambda_1(t_1)\phi\frac{\kappa(r_1-u_m)-1}{I_1(t_f)+\kappa\phi}) > 0$. Therefore, if $I_2(t_f) > I_1(t_f)$ then $t_1 = t_0$, $u^* \equiv 0$ and $I_2(t_0) > I_1(t_0)$.

Case 2: Suppose $I_1(t_f) > I_2(t_f)$. Then

$$\dot{\mathrm{SF}}(t_f) = (I_2(t_f) - I_1(t_f))(1 + \lambda_1(t_f)\phi\frac{\kappa(r_1 - u_m) - 1}{I_2(t_f) + \kappa\phi}) = (I_2(t_f) - I_1(t_f) < 0.$$

Therefore, $u = u_m$ on $J_f = (t_1, t_f]$ and by Claim 5.4 we have $I_1(t_1) > I_2(t_1)$. Suppose $t_1 > t_0$, then we have that $\dot{\mathrm{SF}}(t_1) > 0$. But this is not possible since $\dot{\mathrm{SF}}(t_1) = (I_2(t_1) - I_1(t_1))(1 + \lambda_1(t_1)\phi\frac{\kappa(r_1-u_m)-1}{I_1(t_f)+\kappa\phi}) < 0$. Therefore, if $I_1(t_f) > I_2(t_f)$ then $t_1 = t_0$, $u^* \equiv u_m$ and $I_1(t_0) > I_2(t_0)$.

Case 3: Suppose $I_1(t_f) = I_2(t_f)$. In this case $\dot{\mathrm{SF}}(t_f) = 0$ and

$$\ddot{\mathrm{SF}}(t_f) = (2u - u_m)(I_1(t_f) + \kappa\phi).$$

First note that at any switch time τ, $\dot{\mathrm{SF}}(\tau) = (I_2 - I_1)(1 + \lambda_1\phi\frac{\kappa(r_1-u_m)-1}{I_1+\kappa\phi})$ and therefore $\mathrm{sign}(\dot{\mathrm{SF}}(\tau)) = \mathrm{sign}(I_2 - I_1)$. Therefore if $u(t_f) = u_m$ then $\ddot{\mathrm{SF}}(t_f) > 0$ and so by Claim 5.4 we have that $u^* \equiv u_m$ and $I_1(t) > I_2(t)$ for all $t < t_f$. Similarly, if $u(t_f) = 0$ then $\ddot{\mathrm{SF}}(t_f) < 0$ and so by Claim 5.4 we have that $u^* \equiv 0$ and $I_2(t) > I_1(t)$ for all $t < t_f$. Finally, if $u(t_f) = \frac{u_m}{2}$ then $\ddot{\mathrm{SF}}(t_f) = 0$. Hence if $u^*(t_f) = \frac{u_m}{2}$ then either

(i) $J_f = [t_0, t_f]$, $u^* = \frac{u_m}{2}$ and $I_1(t_0) = I_2(t_0)$,

(ii) $J_f = (t_1, t_f]$, $u^* = u_m$ on $[t_0, t_1]$ and $I_1(t_0) > I_2(t_0)$ or

(iii) $J_f = (t_1, t_f]$, $u^* = 0$ on $[t_0, t_1]$ and $I_1(t_0) < I_2(t_0)$.

□

5.4.3 Towards an Optimal Treatment Strategy

Although a complete solution to Problem 5.2 remains elusive, a particularly interesting observation is that the optimal treatment strategy for model (5.2) has the same form as the optimal treatment strategy for model (5.11) when the two resistant strains have identical characteristics (i.e., compare Theorem 5.1 and Theorem 5.2). Essentially, a single drug should be used until both strains are present in equal amounts and then a mixed strategy should be used to ensure that the strains remain in equal proportions. To gain some insight into why the optimal control has this form for these special cases and also why this form may not be optimal in general, it is useful to examine the equilibria of the system. We begin with some definitions.

Definition. For any constant value $u \equiv u_c$ model (5.11) has the globally stable equilibrium

$$(I_{1eq}(u_c), I_{2eq}(u_c)) = \left(\frac{-\phi[1+\kappa(u_m - u_c)]}{r_1 - u_c}, \frac{-\phi[1+\kappa u_c]}{r_2 - u_m + u_c} \right).$$

The **optimal equilibrium** is obtained by choosing $u \equiv \tilde{u}$ such that

$$I_{1eq}(\tilde{u}) + I_{2eq}(\tilde{u}) = \min_{u_c \in [0, u_m]} \{I_{1eq}(u_c) + I_{2eq}(u_c)\}.$$

The constant $\tilde{u}$ is called the **optimal equilibrium control**.

The motivation for defining the optimal equilibrium is as follows. Under the assumption that the treatment strategy eventually becomes constant then the system will asymptotically approach an equilibrium. Therefore, since the cost function in Problem 5.2 is $\int_{t_0}^{t_f} I_1 + I_2 \mathrm{d}t$ we see that for sufficiently large t_f the cost is dominated by $I_{1eq} + I_{2eq}$ (under the assumption that u is eventually constant). Hence minimizing $I_{1eq} + I_{2eq}$ becomes increasingly important as t_f becomes increasingly large.

Definition. Let $(I_{1eq}(\tilde{u}), I_{2eq}(\tilde{u}))$ be the optimal equilibrium.
Then $\hat{r} = \frac{I_{2eq}(\tilde{u})}{I_{1eq}(\tilde{u})}$ is the **optimal ratio**. Suppose there exists a $t_1 \in [t_0, t_f]$ such that $\frac{I_2(t_1)}{I_1(t_1)} = \hat{r}$. Then $\hat{u}$ is an **optimal ratio control** if setting $u(t) = \hat{u}(t)$ for all $t \geq t_1$ implies that $\frac{I_2(t)}{I_1(t)} = \hat{r}$ for all $t \geq t_1$

In what follows the notation $\tilde{u}$ will denote the optimal equilibrium control (note that $\tilde{u}$ is necessarily constant) and the notation $\hat{u}$ will denote an optimal ratio control (note that $\hat{u}$ is not necessarily constant). Furthermore, notice that if $\kappa = 0$ then model (5.11) is equivalent to model (5.2). Therefore, these definitions also apply to model (5.2).

Observation 5.1. *For model* (5.2) *the optimal equilibrium* $(I_{1eq}(\tilde{u}), I_{2eq}(\tilde{u}))$ *satisfies* $\frac{I_{2eq}(\tilde{u})}{I_{1eq}(\tilde{u})} = 1$. *Furthermore* $\tilde{u} \equiv \hat{u}$.

Proof. For model (5.2) we have that $(I_{1eq}(u), I_{2eq}(u)) = (\frac{-\phi}{r_1-u}, \frac{-\phi}{r_2-u_m+u})$. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}u}(I_{1eq} + I_{2eq}) = \frac{-\phi}{(r_1-u)^2} + \frac{\phi}{(r_2-u_m+u)^2},$$

and so, $\tilde{u} = \frac{r_1 - r_2 + u_m}{2}$ minimizes $I_{1eq} + I_{2eq}$. Furthermore,

$$(I_{1eq}(\tilde{u}), I_{2eq}(\tilde{u})) = \left(\frac{-2\phi}{r_1 + r_2 - u_m}, \frac{-2\phi}{r_1 + r_2 - u_m}\right)$$

and therefore $\frac{I_{2eq}(\tilde{u})}{I_{1eq}(\tilde{u})} = 1$.

To see that $\tilde{u} = \hat{u}$ consider the following: $\frac{I_2}{I_1} \equiv \hat{r} = 1$ implies that $\dot{I}_1 = \dot{I}_2$. This holds only if $\hat{u} = \frac{r_1 - r_2 + u_m}{2} = \tilde{u}$. □

Observation 5.2. *Assume $r_1 = r_2$. Then for model (5.11) the optimal equilibrium $(I_{1eq}(\tilde{u}), I_{2eq}(\tilde{u}))$ satisfies $\frac{I_{2eq}(\tilde{u})}{I_{1eq}(\tilde{u})} = 1$. Furthermore $\tilde{u} \equiv \hat{u}$.*

Proof. For model (5.11) we have that $(I_{1eq}(u), I_{2eq}(u)) = (\frac{-[1+\kappa(u_m - u)]\phi}{r_1 - u}, \frac{-[1+\kappa u]\phi}{r_2 - u_m + u})$. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}u}(I_{1eq} + I_{2eq}) = \frac{[\kappa(r_1 - u_m) - 1]\phi}{(r_1 - u)^2} - \frac{[\kappa(r_2 - u_m) - 1]\phi}{(r_2 - u_m + u)^2},$$

and so, if $r_1 = r_2$ then $\tilde{u} = \frac{u_m}{2}$ minimizes $I_{1eq} + I_{2eq}$. Furthermore,

$$(I_{1eq}(\tilde{u}), I_{2eq}(\tilde{u})) = \left(\frac{-[2 + \kappa u_m]\phi}{2r_1 - u_m}, \frac{-[2 + \kappa u_m]\phi}{2r_1 - u_m}\right)$$

and therefore $\frac{I_{2eq}(\tilde{u})}{I_{1eq}(\tilde{u})} = 1$.

To see that $\tilde{u} = \hat{u}$ consider the following: $\frac{I_2}{I_1} \equiv \hat{r} = 1$ implies that $\dot{I}_1 = \dot{I}_2$. This holds only if $\hat{u} = \frac{r_1 - r_2 + u_m}{2} = \frac{u_m}{2} = \tilde{u}$. □

Using Observations 5.1 and 5.2, Theorems 5.1 and 5.2 can be rephrased in terms of obtaining optimal equilibria:

Corollary 5.1. *The optimal treatment strategy for Problem 5.1 is to target the most* *prevalent strain with maximal effort until the ratio of the two strains reaches the optimal ratio and then to switch to the optimal ratio control.*

Corollary 5.2. *If $r_1 = r_2$ then the optimal treatment strategy for Problem 5.2 is to target the most prevalent strain with maximal effort until the ratio of the two strains reaches the optimal ratio and then to switch to the optimal ratio control.*

In general, the situation is more complicated. If $r_1 \neq r_2$ in model (5.11) then as demonstrated in Figure 5.2 the optimal ratio is is not necessarily 1. In fact, we can show that if $I_{1eq}(\tilde{u}) < I_{2eq}(\tilde{u})$ then it is not optimal for the system to be in equilibrium.

Claim 5.5. *Let*

$$(I_{1eq}(\tilde{u}), I_{2eq}(\tilde{u})) = \left(\frac{-\phi[1+\kappa(u_m - \tilde{u})]}{r_1 - \tilde{u}}, \frac{-\phi[1+\kappa\tilde{u}]}{r_2 - u_m + \tilde{u}} \right)$$

be the optimal equilibrium. If

(i) $r_2 - u_m < r_1 - u_m < r_2 < r_1 < 0$,

(ii) $I_{1eq}(\tilde{u}) < I_{2eq}(\tilde{u})$ *and*

(iii) $(I_1(t_0), I_2(t_0)) = (I_{1eq}(\tilde{u}), I_{2eq}(\tilde{u}))$

then it is not optimal for the system to remain at the optimal equilibrium.

Proof. We will prove this claim by designing a control strategy that results in a lower cost than $\int_{t_0}^{t_f} (I_{1eq}(\tilde{u}) + I_{2eq}(\tilde{u}))\mathrm{d}t$.

Proposed Control Strategy:

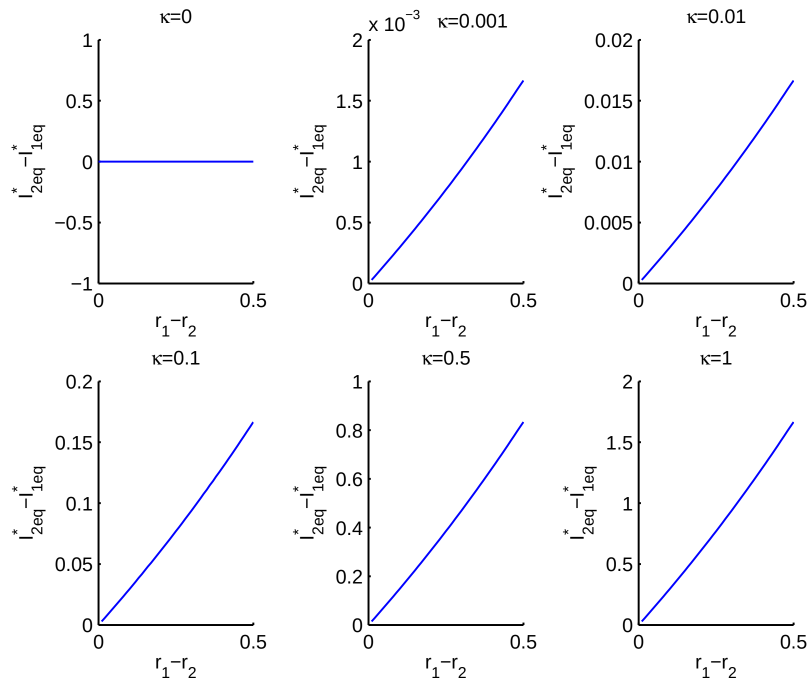

Figure 5.2: Dependence of optimal equilibria on $r_1 - r_2$ and κ. Each panel plots the difference between the optimal equilibria $(I_{2eq}(\tilde{u}) - I_{1eq}(\tilde{u}))$ as a function of the difference between the two strains $(r_1 - r_2)$ for a different value of κ. Notice that if $\kappa > 0$ and $r_1 - r_2 > 0$ then $I_{2eq}(\tilde{u}) - I_{1eq}(\tilde{u}) > 0$. Furthermore notice that $I_{2eq}(\tilde{u}) - I_{1eq}(\tilde{u})$ increases as κ increases and $r_1 - r_2$ increases. This figure suggests that if $\kappa > 0$ and $r_1 - r_2 > 0$ then $I_{2eq}(\tilde{u}) > I_{1eq}(\tilde{u})$. Parameter values are $r_2 = -1.5$, $\phi = 10$ and $u_m = 0.5$.

Assume conditions (i)-(iii) of Claim 5.5 hold. Then define the bang-bang control u_p such that (1) u_p switches to $u = 0$ whenever $I_1 + I_2 = I_{1eq}(\tilde{u}) + I_{2eq}(\tilde{u})$ and (2) u_p switches to $u = u_m$ whenever either $I_1 = I_2$ or $\dot{I}_1 + \dot{I}_2 = 0$.

Remark 5.2. Under the assumption that $I_1(t_0) = I_{1eq}(\tilde{u}) < I_{2eq}(\tilde{u}) = I_2(t_0)$, the control u_p has been designed to drive I_1 and I_2 towards each other ($u_p = 0$) until either (1) the integrand of the cost function is no longer decreasing or (2) the two states are equal to each other (either $\dot{I}_1 + \dot{I}_2 = 0$ or $I_1 = I_2$). Then the control drives the states away from each other (u_p switches to u_m), and continues to do so until the instantaneous cost is equal to the instantaneous cost at equilibrium ($I_1 + I_2 = I_{1eq}(\tilde{u}) + I_{2eq}(\tilde{u})$). At this point the control switches back to $u_p = 0$ and drives the states towards each other again (we will show that this will cause the instantaneous cost to decrease). This cycle continues until $t = t_f$.

We will use the following claim to prove Claim 5.5.

Claim 5.6. *Define $D = I_1 + \frac{r_2 - u_m}{r_1} I_2 + \phi\frac{2+\kappa u_m}{r_1}$ and let $\bar{t}$ be any time such that (i) $D(\bar{t}) = 0$ and (ii) $I_{2eq}(u_m) > I_2(\bar{t}) \geq I_1(\bar{t}) \geq I_{1eq}(u_m)$. Fix $\tilde{t} > \bar{t}$. Then if (i) $u(t) = u_m$ for all $t \in [\bar{t}, \tilde{t})$ and (ii) $u(\tilde{t}) = 0$ then $\dot{I}_1(\tilde{t}) + \dot{I}_2(\tilde{t}) < 0$.*

Proof. First notice that if $u(t) = 0$ then $\dot{I}_1(t) + \dot{I}_2(t) = 0$ if and only if

$$I_1(t) = -\frac{r_2 - u_m}{r_1} I_2(t) - \phi\frac{2 + \kappa u_m}{r_1}.$$

That is, $\dot{I}_1(t) + \dot{I}_2(t) = 0$ if and only if $D(t) = 0$. If $\bar{t}$ is defined as in Claim 5.6 then

$$\begin{aligned}
\left.\frac{\mathrm{d}D}{\mathrm{d}t}\right|_{t=\bar{t}} &= \dot{I}_1 + \frac{r_2 - u_m}{r_1}\dot{I}_2, \\
&= (r_1 - u_m)I_1 + \phi + \frac{r_2 - u_m}{r_1} r_2 I_2 + \frac{r_2 - u_m}{r_1}(1 + \kappa u_m)\phi, \\
&= (r_2 - r_1 + u_m)(r_2 - u_m)\frac{I_2}{r_1} + \frac{\phi}{r_1}[(r_2 - r_1 + u_m) + \kappa u_m (r_2 - r_1)].
\end{aligned}$$

Where the third equality was obtained using the fact that $D(\bar{t}) = 0$. Now,

$$\begin{aligned}
\left.\frac{\mathrm{d}D}{\mathrm{d}t}\right|_{t=\bar{t}} &= (r_2 - r_1 + u_m)(r_2 - u_m)\frac{I_2}{r_1} + \frac{\phi}{r_1}[(r_2 - r_1 + u_m) + \kappa u_m (r_2 - r_1)], \\
&= [r_2(r_2 - r_1) + u_m(r_1 - u_m)]\frac{I_2}{r_1} + \frac{\phi}{r_1}[(r_2 - r_1 + u_m) + \kappa u_m (r_2 - r_1)], \\
&= \frac{r_2 - r_1}{r_1}(r_2 I_2 + \phi(1 + \kappa u_m)) + \frac{u_m}{r_1}((r_1 - u_m)I_2 + \phi), \\
&= \frac{r_2 - r_1}{r_1}\dot{I}_2(\bar{t}) + \frac{u_m}{r_1}((r_1 - u_m)I_2(\bar{t}) + \phi). \qquad (5.13)
\end{aligned}$$

Now since $I_{2eq}(u_m) > I_2(\bar{t})$ we have that $\dot{I}_2(\bar{t}) > 0$ and therefore the first term of Equation (5.13) is positive. Also note that $(r_1 - u_m)I_2 + \phi \leq (r_1 - u_m)I_1 + \phi = \dot{I}_1(\bar{t}) \leq 0$ (since $I_2(\bar{t}) \geq I_1(\bar{t}) \geq I_{1eq}(u_m)$) and therefore the second term of relation (5.13) is non-negative. This implies that $\left.\frac{\mathrm{d}D}{\mathrm{d}t}\right|_{t=\bar{t}} > 0$.

This also implies that $D(t) > 0$ for all $t \in (\bar{t}, \tilde{t}]$. This is because (i) $I_{2eq}(u_m) > I_2(\bar{t}) \geq I_1(\bar{t}) \geq I_{1eq}(u_m)$ and (ii) $u(t) = u_m$ for all $t \in [\bar{t}, \tilde{t})$ guarantee that

$$I_{2eq}(u_m) > I_2(t) \geq I_1(t) \geq I_{1eq}(u_m)$$

for all $t \in [\bar{t}, \tilde{t}]$. In other words, suppose there exist times $t_i \in (\bar{t}, \tilde{t}]$ such that $D(t_i) = 0$ and let t_1 be the smallest such time. Then by Claim 5.6, $\left.\frac{\mathrm{d}D}{\mathrm{d}t}\right|_{t=t_1} > 0$, but by continuity $\frac{\mathrm{d}D}{\mathrm{d}t} > 0$ on a neighbourhood of t_1 which is a contradiction.

To conclude the proof notice that if $u = 0$ then $\dot{I}_1 + \dot{I}_2 < 0$ if and only if $D > 0$. □

Claim 5.6 guarantees that if t_1 and t_2 are two consecutive switch times of u_p such that

(i) u_p switches from 0 to u_m at t_1 and

(ii) $D(t_1) \geq 0$

then switching from $u_p = u_m$ to $u_p = 0$ at $t = t_2$ will cause $I_1 + I_2$ to decrease.

Therefore, it only remains to show that if $u_p(t_0) = 0$ then $\dot{I}_1(t_0) + \dot{I}_1(t_0) < 0$. Consider the following,

$$\begin{aligned} \dot{I}_1(t_0) + \dot{I}_2(t_0) &= r_1 I_{1eq}(\tilde{u}) + (1 + \kappa u_m)\phi + (r_2 - u_m) I_{2eq}(\tilde{u}) + \phi, \\ &= \left[-\frac{r_1}{r_1 - \tilde{u}}[1 + \kappa(u_m - \tilde{u})] - \frac{(r_2 - u_m)}{r_2 - u_m + \tilde{u}}[1 + \kappa\tilde{u}]) + (2 + \kappa u_m) \right] \phi. \end{aligned} \tag{5.14}$$

Therefore, $\dot{I}_1(t_0) + \dot{I}_2(t_0) < 0$ if and only if

$$\tilde{u} > \frac{\kappa u_m(r_2 - u_m) - (r_1 - r_2 + u_m)}{\kappa(r_1 - u_m) - 1 + \kappa(r_2 - u_m) - 1}. \tag{5.15}$$

But, from the expressions for I_{1eq} and I_{2eq} we see that $I_{2eq}(\tilde{u}) > I_{1eq}(\tilde{u})$ if and only if relation (5.15) holds. Therefore, since we began by assuming that $I_{2eq}(\tilde{u}) > I_{1eq}(\tilde{u})$ we can conclude that $\dot{I}_1(t_0) + \dot{I}_2(t_0) < 0$. □

5.5 Discussion

We began by reviewing a study of a hospital infection with three strains (one strain susceptible to two drugs and the other two strains susceptible to only one drug)[14]. This study compared the effects of two different treatment strategies (a mixing strategy and a cycling strategy) under the assumption that there is no de novo resistance. Motivated by this study, we analyzed two different models to consider (i) the effect of a wider range of treatment strategies and (ii) the effect of de novo resistance. For the models we studied we proved that if either (i) there is no de novo resistance or (ii) the two resistant strains have identical characteristics then the optimal treatment strategy is not a constant mixing strategy. Specifically, the treatment strategy that minimizes the total infectious burden is the one that targets the most prevalent strain until both strains are equally prevalent and then switches to a mixed strategy that maintains the two strains in equal proportion. Furthermore, this indicates that de novo resistance changes the form of the optimal treatment strategy only when the two resistant strains have different characteristics.

If treatment can cause resistance and the two resistant strains have different characteristics then it is more difficult to characterize the optimal treatment strategy. We were, however, able to use the solutions for the two simpler cases to gain insight into the more general case. We showed that the fact that the two simpler cases have the same optimal treatment strategy can be explained (at least intuitively) by the fact that the optimal equilibria for the two strains are identical (recall that the optimal equilibria are the equilibrium values of I_1 and I_2 that minimizes $I_1 + I_2$). We showed that, for the simpler cases, if the system is at the optimal equilibrium then it is optimal to remain at that equilibrium. Conversely, we showed that in the general case (provided $I_{1eq}(\tilde{u}) < I_{2eq}(\tilde{u})$) it is not optimal to remain at the optimal equilibrium. We proved this by designing a control that keeps the system out of equilibrium and results in a total cost that is less than $\int_{t_0}^{t_f} I_{1eq}(\tilde{u}) + I_{2eq}(\tilde{u})\mathrm{d}t$. This suggests that when there is de novo resistance and non-identical strains, it is in fact preferable to use a treatment strategy that keeps the system out of equilibrium. In other words, a constant mixing strategy is not optimal.

Although this is only a preliminary study it does raise some important points. Among these is the possibility that de novo resistance and differing strain characteristics can drastically alter the optimal treatment strategy. Furthermore, the extent to which the optimal treatment strategy outperforms the optimal constant mixing strategy may depend on the difference between the two optimal equilibria (i.e., $I_{2eq}(\tilde{u}) - I_{1eq}(\tilde{u})$).

Chapter 6

Summary

In this thesis, Pontryagin's Maximum principle was used to determine optimal intervention strategies to mitigate the spread of infectious diseases. A number of different infectious disease models were analyzed and optimal vaccination, isolation and treatment strategies were discussed.

We began by discussing a very simple treatment model and showed that under a number of simplifying assumptions the effect of treatment can be described using simple geometric arguments. Elaborating on this model we also considered the effects of limited resources and de novo resistance.

In Chapter 3 we addressed the question of minimizing outbreak size under the assumption that there are limited resources. We considered an isolation-only model, a vaccination-only model and a mixed isolation-vaccination model. For the isolation model we showed that, provided there are sufficient resources, it is optimal to isolate with maximal effort for the entire outbreak. Furthermore, if there are not sufficient resources then any isolation strategy that uses all of the available resources is optimal. For the vaccination model, we showed that it is optimal to start vaccinating as soon as possible and to continue vaccinating with maximal effort until either the resources run out or the outbreak ends. Finally for the mixed isolation-vaccination model we proved that it is optimal to begin vaccinating and isolating as soon as possible and to continue vaccinating and isolating with maximal effort until either (i) the isolation resources run out, (ii) the vaccination resources run out or (iii) the outbreak ends. If the isolation resources run out first then it is best to continue vaccinating with maximal effort. On the other hand, if the vaccination resources run out first and there are insufficient isolation resources to continue isolating with maximal effort then any isolation strategy that uses the remainder of the resources is optimal. Thus the optimal mixed isolation-vaccination strategy is not simply a combination of the optimal isolation-only and vaccination-only strategies.

In Chapter 4 we addressed the optimal use of a single antiviral drug to minimize the total outbreak size, under the assumption that treatment causes de novo resistance. We proved that, provided treating immediately cannot avert an outbreak then it is in fact optimal to delay the onset of treatment. We provided an upper bound for the optimal amount of delay in terms of the number of susceptibles and also derived a relation under which delay is never optimal. Furthermore we considered a more detailed model and showed that these general principles seem to still hold.

In Chapter 5 we considered how to use two antiviral drugs to minimize the total infectious burden. We assumed that there are two different strains and each strain is effectively treated by only one drug. If either (1) the two resistant strains have identical properties or (2) treatment does not cause de novo resistance then it is optimal to target the most prevalent strain until both strains are equally prevalent and then to use a constant mixed control to ensure that the two strains remain present in equal amounts. Thus a constant mixed strategy is optimal only when both strains are present in equal amounts. We also considered the general case when the two strains have different characteristics and treatment does cause de novo resistance. In this case, we showed that a constant mixed strategy is not even optimal when the two strains are at the optimal equilibrium. We proved this by designing a control strategy that resulted in a lower cost than the optimal constant mixed strategy.

Bibliography

[1] A. Abakuks. *Optimal Policies for Epidemics.* PhD thesis, University of Sussex, 1972.

[2] A. Abakuks. An optimal isolation policy for an epidemic. *Journal of Applied Probability*, 10(2):247–262, 1973.

[3] A. Abakuks. Optimal immunization policies for epidemics. *Advances in Applied Probability*, 6(3):494–511, 1974.

[4] A. A. Agrachev and Y. L. Sachkov. *Control Theory from the Geometric Viewpoint, Encyclopedia of Mathematical Sciences vol. 87.* Springer-Verlag, New York-Heidelberg-Berlin, 2004.

[5] M. E. Alexander, C. S. Bowman, Z. Feng, M. Gardam, S. M. Moghadas, G. Rst, J. Wu, and P. Yan. Emergence of drug resistance: Implications for antiviral control of pandemic inuenza. *Proc. R. Soc. B*, 274(1619):1675–1684, 2007.

[6] M. E. Alexander, S. M. Moghadas, G. Rost, and J. Wu. A delay differential model for pandemic influenza with antiviral treatment. *Bulletin of Mathematical Biology*, 70(2):382–397, 2008.

[7] R. M. Anderson and R. M. May. *Infectious Diseases of Humans - Dynamics and Control*. Oxford Science Publications, Oxford, 1991.

[8] R. Antia, A. Handel, and I. Longini. What is the best control strategy for multiple infectious disease outbreaks? *Proceedings of the Royal Society B*, 274:833–837, 2007.

[9] N. Arinaminpathy and A. R. McLean. Antiviral treatment for the control of pandemic influenza: Some logistical constraints. *Journal of the Royal Society Interface*, 5(22):545–553, 2008.

[10] J. Arino, F. Brauer, P. van den Driessche, J. Watmough, and J. Wu. Simple models for containment of a pandemic. *Journal of the Royal Society Interface*, 3(8):382–397, 2006.

[11] H. Behncke. Optimal control of deterministic epidemics. *Optimal Control Applications and Methods*, 21:269–285, 2000.

[12] C. T. Bergstrom, M. Lo, and M. Lipsitch. Ecological theory suggests that antimicrobial cycling will not reduce antimicrobial resistance in hospitals. *Proceedings of the National Academy of Sciences*, 101(36):13285, 2004.

[13] L. D. Berkovitz. *Optimal Control Theory, Applied Mathematical Sciences vol. 12*. Springer-Verlag, New York-Heidelberg-Berlin, 1974.

[14] S. Bonhoeffer, M. Lipsitch, and B. R. Levin. Evaluating treatment protocols to prevent antibiotic resistance. *Proceedings of the National Academy of Sciences*, 94(22):12106, 1997.

[15] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproductive ratio r_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382, 1990.

[16] N. M. Ferguson, D. A. T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437, 2005.

[17] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442, 2006.

[18] N. M. Ferguson, S. Mallett, H. Jackson, N. Roberts, and P. Ward. A population-dynamic model for evaluating the potential spread of drug-resistant influenza virus infections during community-based use of antivirals. *J. Antimicrob. Chemother.*, 51(4):977–990, 2003.

[19] W. H. Fleming and R. W. Rishel. *Deterministic and Stochastic Optimal Control*. Springer, 1982.

[20] T. C. Germann, K. Kadau, I. M. Longini, and C. A. Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, 2006.

[21] D. Greenhalgh. Some results on optimal control applied to epidemics. *Mathematical Biosciences*, 88(2):125–158, 1988.

[22] A. Handel, I. M. Longini, and R. Antia. Antiviral resistance and the control

of pandemic influenza: The roles of stochasticity, evolution and model details. *Journal of Theoretical Biology*, 256(1):117–125, 2009.

[23] E. Hansen and T. Day. Optimal control of epidemics with limited resources. *Journal of Mathematical Biology*, 2010, DOI: 10.1007/s00285-010-0341-0.

[24] E. Hansen and T. Day. Optimal antiviral treatment strategies and the effects of resistance. *Proceedings of the Royal Society B: Biological Sciences*, 2010, DOI: 10.1098/rspb.2010.1469.

[25] H. W. Hethcote. The mathematics of infectous diseases. *SIAM Review*, 42(4):599–653, 2000.

[26] M. I. Kamien and N. L. Schwartz. *Dynamic Optimization. The Calculus of Variations and Optimal Control in Economics and Management, Advanced Textbooks in Economics, Vol. 31.* Elsevier, North Holland, 2003.

[27] S. Lenhart and J. Workman. *Optimal Control Applied to Biological Models.* Chapman and Hall, Boca Raton Fl, 2007.

[28] M. Lipsitch, T. Cohen, M. Murray, and B. R. Levin. Antiviral resistance and the control of pandemic influenza. *PLoS Med*, 4(1):0111–0121, 2007.

[29] Jr. Longini, I. M., A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. T. Cummings, and M. E. Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[30] J. M. McCaw, J. G. Wood, C. T. McCaw, and J. McVernon. Impact of emerging antiviral drug resistance on influenza containment and spread: Influence of

subclinical infection and strategic use of a stockpile containing one or two drugs. *PLoS ONE*, 2008.

[31] S. Merler, M. Ajelli, and C. Rizzo. Age-prioritized use of antivirals during an influenza pandemic. *BMC Infectious Diseases*, 9(1):117, 2009.

[32] S. M. Moghadas. Management of drug resistance in the population: Influenza as a case study. *Proceedings of the Royal Society B: Biological Sciences*, 275(1639):1163, 2008.

[33] S. M. Moghadas, C. S. Bowman, G. Rost, and J. Wu. Population-wide emergence of antiviral resistance during pandemic influenza. *PLoS ONE*, 2008.

[34] P. M. Oostvogel, H. G. A. M. van der Avoort, M. N. Mulders, A. M. van Loon, M. A. E. Conyn van Spaendonck, H. C. Rmke, G. van Steenis, and J. K. van Wijngaarden. Poliomyelitis outbreak in an unvaccinated community in the netherlands, 1992-93. *The Lancet*, 344(8923):665–670, 1994.

[35] L. S. Pontryagin, V. G. Boltyanski, R. V. Gamkrelidze, and E. F. Mishchenko. *The Mathematical Theory of Optimal Processes, International Series of Monographs in Pure and Applied Mathematics vol. 55.* The Macmillan Company, New York, 1964.

[36] Z. Qiu and Z. Feng. Transmission dynamics of an influenza model with vaccination and antiviral treatment. *Bulletin of Mathematical Biology*, 72(1):1–33, 2010.

[37] K. H. Wickwire R. Morton. On the optimal control of a deterministic epidemic. *Advances in Applied Probability*, 6(4):622–635, 1974.

[38] R. R. Regoes and S. Bonhoeffer. Emergence of drug-resistant influenza virus: Population dynamical considerations. *Science*, 312(5772):389–391, 2006.

[39] N. I. Stilianakis, A. S. Perelson, and F. G. Hayden. Emergence of drug resistance during an influenza epidemic: Insights from a mathematical model. *The Journal of Infectious Diseases*, 177(4):863–873, 1998.

[40] D. M. Weinstock and G. Zuccotti. The evolution of influenza resistance and treatment. *The Journal of the American Medical Association*, 301:1066–1069, 2009.

[41] K. H. Wickwire. Optimal isolation policies for deterministic and stochastic epidemics. *Mathematical Biosciences*, 26(3-4):325–346, 1975.

[42] K. H. Wickwire. Mathematical models for the control of pests and infectious diseases: A survey. *Theoretical Population Biology*, 11(2):182–238, 1977.

[43] J. T. Wu, G. M. Leung, M. Lipsitch, B. S. Cooper, and S. Riley. Hedging against antiviral resistance during the next influenza pandemic using small stockpiles of an alternative chemotherapy. *PLoS Med*, 2009.

Appendix A

Existence Proofs

Here we explicitly prove the existence of an optimal solution for the various optimal control problems we have discussed. Theorem 1.2 from Chapter 1 can be used to prove that a solution exists for the optimal control problems considered in Chapters 3, 4 and 5.

The optimal control problems in Chapter 3 have the following form:

General Form A.1. Fix constants S_0, I_0, $u_{\max}$, β, $\mu > 0$, $w_{\max}$, $z_{\max} \geq 0$ and $I_0 > r > 0$. Define the following system of differential equations

$$\begin{aligned} \dot{S} &= -\beta S I - u_v S, \\ \dot{I} &= \beta S I - (\mu + u_i) I, \\ \dot{w} &= u_i I, \\ \dot{z} &= u_v S, \end{aligned} \tag{A.1}$$

with initial conditions $S(t_0) = S_0 > 0$, $I(t_0) = I_0 > 0$, $w(t_0) = z(t_0) = 0$. Find the measurable functions $u_v : [t_0, t_f] \to [0, u_{\max}]$ and $u_i : [t_0, t_f] \to [0, u_{\max}]$ that minimize the cost function

$$J(u_v, u_i) = \int_{t_0}^{t_f} \beta S I \; \mathrm{d}t,$$

subject to $t_f = \min\{t \mid I(t) = r\}$, $w(t_f) \leq w_{\max}$ and $z(t_f) \leq z_{\max}$.

Claim A.1. *There exists an optimal solution for General Problem A.1.*

Proof. All conditions of Theorem 1.2 hold trivially except for condition (v). We will now show that there exists a $\bar{t} > t_0$ such that for any choice of u_i and u_v we have $t_f \leq \bar{t}$. This shows that the set of allowable boundary conditions $\mathcal{B}$ can be defined as a compact set and therefore Condition (v) holds.

Fix small $\epsilon > 0$. Then for model (A.1) we have the following two observations:

(i) $\dot{S} = -(\beta I + u_v)S \leq -\beta r S$,

(ii) if $S \leq \frac{\mu}{\beta} - \epsilon$ then $\dot{I} = (\beta S - \mu - u_i)I \leq (\beta \frac{\mu}{\beta} - \beta\epsilon - \mu)I = -\beta\epsilon I$.

Case 1: Assume $S_0 > \frac{\mu}{\beta} - \epsilon$. Define $\tilde{S} : \mathbb{R} \to \mathbb{R}$ by $\tilde{S}(t_0) = S_0$ and $\dot{\tilde{S}} = -\beta r \tilde{S}$. Then there exists a time t_1 such that $\tilde{S}(t_1) = \frac{\mu}{\beta} - \epsilon$. Now for any particular choice of u_i and u_s either (1) $S(t_f) > \frac{\mu}{\beta} - \epsilon$ or (2) $S(t_1) < \tilde{S}(t_1) = \frac{\mu}{\beta} - \epsilon$. If (1) is true then $t_f < t_1$. If on the other hand, (2) is true then t_1 is an upper bound for the amount of time it takes S to decrease to $\frac{\mu}{\beta} - \epsilon$. Now define $\tilde{I} : \mathbb{R} \to \mathbb{R}$ by $\tilde{I}(t_0) = S_0$ and $\dot{\tilde{I}} = -\beta\epsilon\tilde{I}$. Then there exists a time t_2 such that $\tilde{I}(t_2) = r$. By Observation (ii) we see that $I(t_2) < \tilde{I}(t_2)$. Therefore, if we set $\bar{t} = t - 1 + t_2$ then $t_f \leq \bar{t}$.

Case 2: Assume $S_0 \leq \frac{\mu}{\beta} - \epsilon$. In this case if we define t_2 as above and set $\bar{t} = t_2$, then $t_f \leq \bar{t}$.

Therefore, the set of allowable boundary conditions can be defined as

$$\mathcal{B} = t_0 \times S_0 \times I_0 \times 0 \times 0 \times [0, \bar{t}] \times [0, S_0] \times r \times [0, w_{\max}] \times [0, z_{\max}].$$

$\square$

The optimal control problem in Chapter 4 has the following form:

General Form A.2. Fix constants S_0, I_0, $u_{\max}$, β_I, β_R, μ_I, μ_R, $\kappa > 0$ and $I_0 > r > 0$. Define the following system of differential equations

$$\begin{aligned} \dot{S} &= -(\beta_I I + \beta_R R)S, \\ \dot{I} &= \beta_I IS - (\mu_I + u)I, \\ \dot{R} &= \beta_I RS + \kappa u I - \mu_R R, \end{aligned} \tag{A.2}$$

with initial conditions $S(t_0) = S_0 > 0$, $I(t_0) = I_0 > 0$ and $R(t_0) = 0$. Find the measurable function $u : [t_0, t_f] \to [0, u_{\max}]$ that minimizes the cost function

$$J(u) = \int_{t_0}^{t_f} (\beta_I I + \beta_R R)S \text{ dt} = S_0 - S(t_f),$$

subject to $t_f = \min\{t \mid I(t) + R(t) = r\}$.

Claim A.2. *There exists an optimal solution for General Problem A.2.*

Proof. All conditions of Theorem 1.2 hold trivially except for condition (v). We will show that there exists a constant $B > \inf J$ and a $\bar{t} > t_0$ such that if $J(u) < B$ then $t_f < \bar{t}$. This implies that any u that results in $t_f > \bar{t}$ cannot be optimal and therefore the set of allowable boundary conditions $\mathcal{B}$ can be defined as a compact set (i.e., Condition (v) holds).

Let $\hat{S}_f$ denote the final number of susceptibles when $u \equiv 0$. Since $J(u) = \int_{t_0}^{t_f} (\beta_I I + \beta_R R) S \text{ d}t = S_0 - S(t_f)$, the cost is completely specified by the final number of susceptibles. Now define $\tilde{S} : \mathbb{R} \to \mathbb{R}$ by $\tilde{S} = S_0$ and $\dot{\tilde{S}} = -\min\{\beta_I, \beta_R\} r S$. Then there exists a time $\bar{t}$ such that $\tilde{S}(\bar{t}) = \bar{S}_f$. Furthermore, since

$$\dot{S} = -(\beta_I I + \beta_R R) S \leq -\min\{\beta_I, \beta_R\} r S,$$

whenever $J(u) \leq S_0 - \bar{S}_f$ we have that $t_f \leq \bar{t}$. That is, we can choose

$$\mathcal{B} = t_0 \times S_0 \times I_0 \times 0 \times [0, \bar{t}] \times [0, S_0] \times G,$$

where $G = \{(I, r - I) \mid I \in [0, r]\}$. □

The optimal control problem in Chapter 5 has the following form:

General Form A.3. Fix constants $I_{1,0}$, $I_{2,0}$, $u_{\max}$, $\phi > 0$, $\kappa \geq 0$, r_1, $r_2 < 0$ and $t_f > t_0$. Define the following system of differential equations

$$\begin{aligned} \dot{I}_1 &= (r_1 - u) I_1 + (1 + \kappa(u_{\max} - u))\phi, \\ \dot{I}_2 &= (r_2 - u_{\max} + u) I_2 + (1 + \kappa u)\phi. \end{aligned} \tag{A.3}$$

with initial conditions $I_1(t_0) = I_{1,0}$ and $I_2(t_0) = I_{2,0}$. Find the measurable function

$u : [t_0, t_f] \to [0, u_{\max}]$ that minimizes the cost function

$$J(u) = \int_{t_0}^{t_f} (I_1 + I_2)\ \mathrm{d}t.$$

Claim A.3. *There exists an optimal solution for General Problem A.3.*

Proof. All conditions of Theorem 1.2 hold trivially. $\square$