Research article

A two-strain model of infectious disease spread with asymmetric temporary immunity periods and partial cross-immunity

Matthew D. Johnston*, Bruce Pell and David A. Rubel

Department of Mathematics + Computer Science, Lawrence Technological University, 21000 W 10 Mile Rd, Southfield, MI 48075, USA

* **Correspondence:** Email: mjohnsto1@ltu.edu.

Abstract: We introduce a two-strain model with asymmetric temporary immunity periods and partial cross-immunity. We derive explicit conditions for competitive exclusion and coexistence of the strains depending on the strain-specific basic reproduction numbers, temporary immunity periods, and degree of cross-immunity. The results of our bifurcation analysis suggest that, even when two strains share similar basic reproduction numbers and other epidemiological parameters, a disparity in temporary immunity periods and partial or complete cross-immunity can provide a significant competitive advantage. To analyze the dynamics, we introduce a quasi-steady state reduced model which assumes the original strain remains at its endemic steady state. We completely analyze the resulting reduced planar hybrid switching system using linear stability analysis, planar phase-plane analysis, and the Bendixson-Dulac criterion. We validate both the full and reduced models with COVID-19 incidence data, focusing on the Delta (B.1.617.2), Omicron (B.1.1.529), and Kraken (XBB.1.5) variants. These numerical studies suggest that, while early novel strains of COVID-19 had a tendency toward dramatic takeovers and extinction of ancestral strains, more recent strains have the capacity for co-existence.

Keywords: epidemiology; mathematical model; multi-strain; cross-immunity; reproduction number

1. Introduction

Since the start of the COVID-19 pandemic, compartmental differential equation-based mathematical models have played an instrumental role in informing public health decisions. Early on, modeling efforts focused on forecasting and evaluating the efficacy of non-pharmaceutical interventions such as lockdowns [1, 2], school closures [3, 4], contact tracing [5, 6], and face mask utilization [7, 8]. Following the rapid development of effective vaccines and antiviral treatments, the emphasis shifted to modeling pharmaceutical measures, such as estimating the effects of different vaccine distribution strategies [9, 10], evaluating vaccine efficacy [11, 12], and assessing waning vaccine immunity [13].

Despite the unprecedented global effort to eradicate COVID-19, however, the disease has continued to spread widely and is now widely considered to be endemic in the global population.

A significant factor in this continuing spread has been the emergence of variants, such as Delta (B.1.617.2), Omicron (B.1.1.529) and the Omicron subvariant Kraken (XBB.1.5), which were reported as the most transmissible strains at the time. Transmissibility, however, is only one of several factors, including vaccine-resistance, diagnostic evasion, and cross-immunity, which cumulatively determine whether emerging strains exhibit *competitive exclusion* or *coexistence* with the circulating ancestral strains. Competitive exclusion, where one strain drives the others to extinction, has been observed with the Delta and Omicron waves during the COVID-19 pandemic, where earlier strains were largely eradicated over the course of several months. Indeed, such dramatic takeovers are expected during the beginning of a pandemic by a novel virus as it is rapidly mutating and exploring its evolutionary fitness. As COVID-19 transitions from a pandemic to endemic infection, however, it is anticipated that we will see multiple co-existing strains circulating in a given year, akin to seasonal influenza [14].

Mathematically, it has been common to account for multiple strains in an SIR-type (Susceptible-Infectious-Removed) model of infectious disease spread [15] by dividing the infectious class between several competing and non-overlapping strains. The roots of this research can be found in several areas, including the study of viruses such as influenza, bacterial infections and parasites [16–19]. For many basic SIR-type models extended to multiple strains, the only possibility is competitive exclusion [16, 20–22]. However, there are several mechanisms known to facilitate coexistence, including co-infection, cross-immunity, seasonal variations and age-stratification of infectious [23]. In the context of COVID-19, several studies have demonstrated the capacity of coexistence in certain cases, such as utilizing different force of infection terms [24, 25] and strain-specific vaccination [14], and cross-immunity [13]. Understanding the capacity of a multistrain model to exhibit competitive exclusion or coexistence remains challenging, however, due to the high-dimensionality of the associated models.

In this paper, we introduce a two-strain model which incorporates asymmetric temporary immunity periods and partial cross-immunity. We determine explicit conditions for competitive exclusion and coexistence of the two strains based on the basic reproduction numbers, the temporary immunity period durations and the degree of cross-immunity. To address the challenges of analyzing the dynamics of multistrain models, we further introduce a dimension-reducing modeling framework for analyzing the emergence of new strains of a virus. In our method, we assume that the original strain is endemic in the population and remains at a steady state throughout the evolution of the new strain. This allows the dynamics of the emerging strain to be reduced while closely capturing the dynamics and long-term behavior of the full model. The method is based on the quasi-steady state approximation (QSSA), which is extensively used in the study of biochemical reaction networks to justify the Michaelis-Menten [26, 27] and Hill [28] rate forms. More broadly in mathematical biology, the use of the QSSA is more limited but includes application to a two-strain tuberculosis model [29], parasite-host models [30], cancer therapy [31], vector-borne illnesses [32] and a two-strain dengue fever model [33]. We validate the models by parameter-fitting to COVID-19 incidence data in the United States across multiple strains, including the Delta, Omicron and Kraken variants. The results suggest that differences in the temporary immunity periods and partial cross-immunity are sufficient to account for the dramatic shifts in variant proportions we have seen during the course of the COVID-19 pandemic.

2. Main results

2.1. Basic two-strain model

Consider the following two-strain compartmental model constructed after the classical SIR (Susceptible-Infectious-Removed) model introduced by Kermack and McKendrick in 1927 [15]:
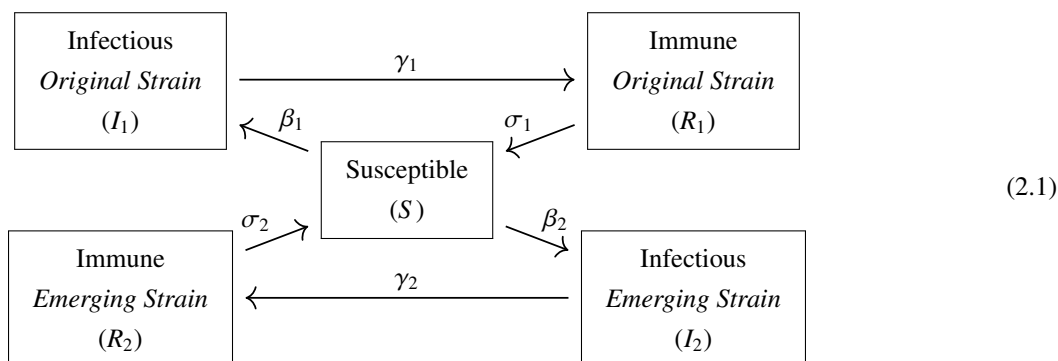
In the model (2.1), we allow each strain to infect a common pool of susceptible individuals at different rates (β_1 and β_2, respectively, for the original and emerging strain). We also allow each strain to exhibit different mean infection periods (γ_1^{-1} and γ_2^{-1}) and different mean temporary immunity periods (σ_1^{-1} and σ_2^{-1}). We do not allow co-infection and assume complete cross-immunity so that individuals catch one strain at a time and completely recover before they are capable of being infected by either strain again. The model (2.1) is a special case of those considered in [21, 22].

The model (2.1) can be corresponded to the following system of ordinary differential equations:

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{S}{N}(\beta_1 I_1 + \beta_2 I_2) + \sigma_1 R_1 + \sigma_2 R_2, & \\[2mm] \dfrac{dI_1}{dt} = \dfrac{\beta_1}{N} S I_1 - \gamma_1 I_1, & \dfrac{dI_2}{dt} = \dfrac{\beta_2}{N} S I_2 - \gamma_2 I_2, \\[2mm] \dfrac{dR_1}{dt} = \gamma_1 I_1 - \sigma_1 R_1, & \dfrac{dR_2}{dt} = \gamma_2 I_2 - \sigma_2 R_2. \end{cases} \quad (2.2)$$

Note that the Eqs (2.2) imply that the total population is constant, i.e., $N = S + I_1 + R_1 + I_2 + R_2$. Setting $S = N - I_1 - R_1 - I_2 - R_2$, we reduce the system to:

$$\begin{cases} \dfrac{dI_1}{dt} = \dfrac{\beta_1}{N}(N - I_1 - R_1 - I_2 - R_2)I_1 - \gamma_1 I_1, & \dfrac{dI_2}{dt} = \dfrac{\beta_2}{N}(N - I_1 - R_1 - I_2 - R_2)I_2 - \gamma_2 I_2, \\[2mm] \dfrac{dR_1}{dt} = \gamma_1 I_1 - \sigma_1 R_1, & \dfrac{dR_2}{dt} = \gamma_2 I_2 - \sigma_2 R_2. \end{cases} \quad (2.3)$$

It can be easily computed that the system (2.2) only permits the following steady states, where $\mathbf{x} = (S, I_1, R_1, I_2, R_2)$:

$$\mathbf{x}_0 = (N, 0, 0, 0, 0) \quad (2.4)$$

$$\mathbf{x}_1 = \left(\frac{N\gamma_1}{\beta_1}, \frac{N\sigma_1(\beta_1 - \gamma_1)}{\beta_1(\gamma_1 + \sigma_1)}, \frac{N\gamma_1(\beta_1 - \gamma_1)}{\beta_1(\gamma_1 + \sigma_1)}, 0, 0 \right) \quad (2.5)$$

$$\mathbf{x}_2 = \left(\frac{N\gamma_2}{\beta_2}, 0, 0, \frac{N\sigma_2(\beta_2 - \gamma_2)}{\beta_2(\gamma_2 + \sigma_2)}, \frac{N\gamma_2(\beta_2 - \gamma_2)}{\beta_2(\gamma_2 + \sigma_2)} \right) \tag{2.6}$$

The *disease free steady state* $\mathbf{x}_0$ exists for all parameter values, while the *original strain only steady state* $\mathbf{x}_1$ is physically relevant if and only if $\beta_1 > \gamma_1$ and the *emerging strain only steady state* $\mathbf{x}_2$ is physically relevant if and only if $\beta_2 > \gamma_2$. The system does not have the capacity for a *coexistence steady state*, i.e., a steady state $\mathbf{x}_{12}$ with $I_1 > 0$ and $I_2 > 0$.

2.2. Basic reproduction number

The basic reproduction number of a disease, denoted $\mathscr{R}_0$, is one of the most important and well-studied parameters in the study of infectious disease spread. Intuitively, it corresponds to the expected number of secondary infections produced by a single primary infection in a fully susceptible population [34]. Consequently, an infectious disease has the capacity to infiltrate a population if and only if $\mathscr{R}_0 > 1$. For multi-strain models like (2.1), however, we are also interested in the capacity of an individual *strain* to infiltrate a population, either in the absence or presence of other strains. Following the notation of [35], we let the basic reproduction number of strain i in the absence of other strains be denoted by $\mathscr{R}_i$, and the basic reproduction number of strain i in the presence of strain j be denoted by $\mathscr{R}_{ij}$. The threshold $\mathscr{R}_i > 1$ suggests that strain i has the capacity to infiltrate the population in the absence of other strains, while the condition $\mathscr{R}_{ij} > 1$ suggests that strain i has the capacity to infiltrate a population already infected at a steady level with strain j.

For compartmental differential equations models, the next-generation method can be used to calculate the basic reproduction numbers [35–39] (see Appendix A for details). For the basic two-strain model (2.3), the next generation method gives the parameters (see Appendix A.1 for details):

$$\mathscr{R}_1 = \frac{\beta_1}{\gamma_1}, \mathscr{R}_2 = \frac{\beta_2}{\gamma_2}, \mathscr{R}_{12} = \frac{\mathscr{R}_1}{\mathscr{R}_2}, \mathscr{R}_{21} = \frac{\mathscr{R}_2}{\mathscr{R}_1}, \text{ and } \mathscr{R}_0 = \max\{\mathscr{R}_1, \mathscr{R}_2\}.$$

It follows from $\mathscr{R}_i > 1$, $i = 1, 2$, that strain i has the capacity to infiltrate a disease-free population only if $\beta_i > \gamma_i$, and from $\mathscr{R}_0 = \max\{\mathscr{R}_1, \mathscr{R}_2\} > 1$ that the disease will infiltrate the population if at least one strain is able to infiltrate. The condition $\mathscr{R}_{ij} > 1$ suggests the strain i has the capacity to infiltrate a population already infected with strain j if $\mathscr{R}_i > \mathscr{R}_j$, i.e., its basic reproduction number must be strictly higher than that of the other strain.

It is furthermore known from the results of [21, 22] that: (a) the disease free steady state $\mathbf{x}_0$ (2.4) is stable if $\mathscr{R}_1 < 1$ and $\mathscr{R}_2 < 1$; (b) the original strain only steady state $\mathbf{x}_1$ (2.5) is stable if $\mathscr{R}_1 > 1$ and $\mathscr{R}_1 > \mathscr{R}_2$; and (c) the emerging strain only steady state $\mathbf{x}_2$ (2.6) is stable if $\mathscr{R}_2 > 1$ and $\mathscr{R}_2 > \mathscr{R}_1$. Which individual strains survives is only dependent on the relative values of the basic reproduction numbers $\mathscr{R}_1$ and $\mathscr{R}_2$. Ultimately, the more contagious strain will infiltrate the population and eliminate the other strain; coexistence of strains is not permitted in the long-term dynamics.

2.3. Full asymmetric temporary immunity periods and partial cross-immunity model

We now extend the two-strain model (2.1) to include asymmetric temporary immunity periods and partial cross-immunity:

(2.7)
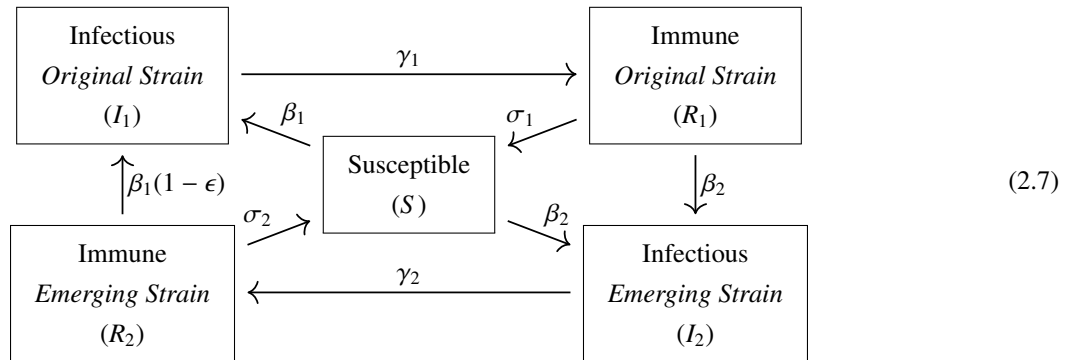
In the model (2.7), we assume individuals in the temporary immunity states R_1 and R_2 can only be infected by the *other* strain. Those recovered from the original strain (R_1) are infected by the emerging strain at the same rate as susceptible individuals (S); however, those recovered from the emerging strain (R_2) are provided with partial protection from catching the original strain. This asymmetric cross-immunity can occur when the antibodies produced from infection by the emerging strain are sufficient protection against the original strain, but not the other way around. The degree of cross-immunity is tuned by the parameter $0 \le \epsilon \le 1$, with $\epsilon = 0$ corresponding to no cross-immunity, and $\epsilon = 1$ corresponding to complete cross-immunity.

Note that we have opted not to include in (2.7) a latency period, which is often incorporated in the form of an exposed class of individuals. The decision to not include a latency period was made for mathematical tractability and to allow us to focus on other key aspects of the disease dynamics, such as transmission rates, immunity periods and the impact of emerging variants. While the latency period is an important component of the disease process, its omission does not significantly alter the general dynamics of the model leading to similar results for the reproduction numbers [22, 39]. However, we note that the latency period plays a role in the understanding of disease transmission and can influence the effectiveness of control measures, such as testing and lockdown protocols.

We utilize the following system of ODEs to model the time evolution of the full model (2.7):

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{S}{N}\left(\beta_1 I_1 + \beta_2 I_2\right) + \sigma_1 R_1 + \sigma_2 R_2, \\[2mm] \dfrac{dI_1}{dt} = \dfrac{\beta_1}{N} I_1(S + (1-\epsilon)R_2) - \gamma_1 I_1, \qquad \dfrac{dI_2}{dt} = \dfrac{\beta_2}{N} I_2(S + R_1) - \gamma_2 I_2, \\[2mm] \dfrac{dR_1}{dt} = \gamma_1 I_1 - \sigma_1 R_1 - \dfrac{\beta_2}{N} R_1 I_2, \qquad \dfrac{dR_2}{dt} = \gamma_2 I_2 - \sigma_2 R_2 - \dfrac{\beta_1}{N}(1-\epsilon)R_2 I_1. \end{cases} \quad (2.8)$$

Substituting $S = N - I_1 - R_1 - I_2 - R_2$ into (2.8), we obtain the following equivalent system of ODEs:

$$\begin{cases} \dfrac{dI_1}{dt} = \dfrac{\beta_1}{N}(N - I_1 - R_1 - I_2 - \epsilon R_2)I_1 - \gamma_1 I_1, \quad \dfrac{dI_2}{dt} = \dfrac{\beta_2}{N}(N - I_1 - I_2 - R_2)I_2 - \gamma_2 I_2, \\[2mm] \dfrac{dR_1}{dt} = \gamma_1 I_1 - \sigma_1 R_1 - \dfrac{\beta_2}{N} R_1 I_2, \qquad \dfrac{dR_2}{dt} = \gamma_2 I_2 - \sigma_2 R_2 - \dfrac{\beta_1}{N}(1-\epsilon)R_2 I_1. \end{cases} \quad (2.9)$$

The analysis of (2.9) is more complicated than that of (2.3). Nevertheless, it can be easily checked that the three steady states of (2.3) from Section 2.1, (2.4)–(2.6), are also steady states of (2.9). The basic reproductive numbers listed below (2.10) can be determined by using the next generation method [35] (see Appendix A.2 for details):

$$\mathscr{R}_1 = \frac{\beta_1}{\gamma_1}, \mathscr{R}_2 = \frac{\beta_2}{\gamma_2}, \mathscr{R}_{12} = \frac{\mathscr{R}_1}{\mathscr{R}_2}\left(\frac{(1-\epsilon)\beta_2 + \epsilon\gamma_2 + \sigma_2}{\gamma_2 + \sigma_2}\right), \mathscr{R}_{21} = \frac{\mathscr{R}_2}{\mathscr{R}_1}\left(\frac{\beta_1 + \sigma_1}{\gamma_1 + \sigma_1}\right), \text{ and } \mathscr{R}_0 = \max\{\mathscr{R}_1, \mathscr{R}_2\}.$$

(2.10)

The condition $\mathscr{R}_{21} > 1$, required for the emerging strain to infiltrate a population already infected at an endemic level with the original strain, can be intuitively interpreted by considering limiting cases. If the original strain has a short temporary immunity period (i.e., $\sigma_1 \to \infty$, $\sigma_1^{-1} \to 0$), the emerging strain needs to be more contagious than the original strain to gain a foothold in the population ($\mathscr{R}_2 > \mathscr{R}_1$). This is due to the emerging strain constantly having to compete with the original strain for new infections. If the original strain has a longer temporary immunity period (i.e., $\sigma_1 \to 0$, $\sigma_1^{-1} \to \infty$), however, the emerging strain only needs to be able to sustain itself in the population on its own to survive ($\mathscr{R}_2 > 1$). This is due to the emerging strain having exclusive ability to infect those who have previously been infected with the original strain.

Table 1. Variables and parameters for the full model (2.9) and reduced model (2.13).

Variable	Units	Description
$S \geq 0$	people	Susceptible individuals
$I_i \geq 0$	people	Infectious individuals (i^{th} strain)
$R_i \geq 0$	people	Temporarily immune individuals (i^{th} strain)
$t \geq 0$	days	Time elapsed
Parameter	**Units**	**Description**
$\beta_i \geq 0$	days^{-1}	Transmission rate (i^{th} strain)
$\gamma_i^{-1} \geq 0$	days	Infectious period (i^{th} strain)
$\sigma_i^{-1} \geq 0$	days	Temporary immunity period (i^{th} strain)
$0 \leq \epsilon \leq 1$	—	Degree of cross-immunity
$\mathscr{R}_0 \geq 0$	—	Basic reproduction number of disease
$\mathscr{R}_i \geq 0$	—	Basic reproduction number of strain i
$\mathscr{R}_{ij} \geq 0$	—	Basic reproduction number of strain i in presence of strain j

Similar intuition holds for the condition $\mathscr{R}_{12} > 1$ required for the original strain to survive as the emerging strain infiltrates the population. If the immunity period of the emerging strain is short (i.e., $\sigma_2 \to \infty$, $\sigma_2^{-1} \to 0$), or the degree of cross-immunity provided by the emerging strain is high (i.e., $\epsilon \to 1$), then the original strain will survive only if it is more contagious than the emerging strain ($\mathscr{R}_1 > \mathscr{R}_2$). If, however, the immunity period of the emerging strain is long (i.e., $\sigma_2 \to 0$, $\sigma_2^{-1} \to \infty$) then the original strain will survive only if it can survive on its own ($\mathscr{R}_1 > 1$).

Unlike the basic model (2.1), the model (2.7) also allows for a co-existence steady state where both strains survive and circulate in the population, so long as specific conditions on the basic reproduction numbers, temporary immunity periods, and degree of cross-immunity are satisfied. This is more formally stated in the following result, which we prove in Appendix B.

Theorem 1. *The full two-strain model with asymmetric temporary immunity periods and partial cross-immunity* (2.8) *has a coexistence steady state (i.e.,* $\mathbf{x}_{12} = (S, I_1, R_1, I_2, R_2)$ *with* $I_1 > 0$ *and* $I_2 > 0$*) if and only if* $\min\{\mathscr{R}_1, \mathscr{R}_2, \mathscr{R}_{12}, \mathscr{R}_{21}\} > 1$*, where* $\mathscr{R}_1, \mathscr{R}_2, \mathscr{R}_{12}$*, and* $\mathscr{R}_{21}$ *are the basic reproduction numbers from* (2.10)*. Furthermore, this coexistence steady state is unique whenever it exists.*

2.4. Reduced asymmetric temporary immunity periods and partial cross-immunity model

Although we are able to determine conditions for the existence of four steady states of (2.9), analyzing the dynamics remains difficult to perform directly due to the nonlinearities, four-dimensional state space and seven undetermined parameters. We are also not able to write down an explicit closed form for the co-existence steady state. This makes linear stability and bifurcation analysis challenging.

To make analysis of the model's dynamics more tractable, we perform a model reduction of (2.9). We are primarily interested in the situation where the original strain has become endemic in the population before the emerging strain arrives. This suggests the modeling assumption that the original strain remains at its endemic steady state throughout the dynamics of the emerging strain. To update the dynamics of the emerging strain, we substitute the steady state values of I_1 and R_1 into the dynamical equations for I_2 and R_2. This produces a system of two differential equations for the emerging strain I_2 and R_2, and two algebraic equations for the original strain I_1 and R_1. Consequently, we assume that the first two equations in (2.9) are at steady state. This gives the following system of equations:

$$\begin{cases} \dfrac{\beta_1}{N}(N - I_1 - R_1 - I_2 - \epsilon R_2)I_1 - \gamma_1 I_1 = 0, \\[2mm] \gamma_1 I_1 - \sigma_1 R_1 - \dfrac{\beta_2}{N}R_1 I_2 = 0. \end{cases} \tag{2.11}$$

Solving (2.11) gives rise to the following function, which tracks the steady state level of I_1 as a function of the infection level of I_2 and R_2:

$$\omega(I_2, R_2) = \begin{cases} \dfrac{(N(\beta_1 - \gamma_1) - \beta_1 I_2 - \beta_1 \epsilon R_2)(\beta_2 I_2 + N\sigma_1)}{\beta_1(\beta_2 I_2 + N(\gamma_1 + \sigma_1))}, & \text{if } I_2 + \epsilon R_2 < N\left(1 - \dfrac{1}{\mathscr{R}_1}\right) \\[4mm] 0, & \text{if } I_2 + \epsilon R_2 \geq N\left(1 - \dfrac{1}{\mathscr{R}_1}\right). \end{cases} \tag{2.12}$$

We substitute the steady state function $I_1 = \omega(I_2, R_2)$ into (2.9) to get the reduced two-strain with asymmetric temporary immunity periods and partial cross-immunity model:

$$\begin{cases} \dfrac{dI_2}{dt} = \dfrac{\beta_2}{N}(N - \omega(I_2, R_2) - I_2 - R_2)I_2 - \gamma_2 I_2, \\[3mm] \dfrac{dR_2}{dt} = \gamma_2 I_2 - \sigma_2 R_2 - \dfrac{\beta_1}{N}(1 - \epsilon)\omega(I_2, R_2)R_2. \end{cases} \tag{2.13}$$

Notice that (2.13) is a planar system which only depends on the emerging strain (I_2 and R_2). The original strain is tracked through the steady state function $I_1 = \omega(I_2, R_2)$ (2.12).

Since $\omega(I_2, R_2)$ is a piecewise-defined function, the system (2.13) is a state-dependent switching system [40]. The dynamics of such systems are more varied than standard dynamical systems but have been increasingly studied in recent years due to their applications in control theory [41–43]. We note, in particular, that the right-hand side of (2.13) is continuous everywhere but not differentiable at

$I_2 + \epsilon R_2 = N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$. The system (2.13) is planar and consequently its dynamics can be analyzed by methods such as linear stability analysis, phase-plane analysis, and the Bendixson-Dulac criterion [44].

The method used to reduce (2.9) to (2.13) mirrors that of the quasi-steady state assumption (QSSA). The QSSA is used when there is a parametrizable time-scale separation between two parts of a process and has been popularly used in biochemistry to justify the Michaelis-Menten [27] and Hill [28] kinetic rate functions. This and other asymptotic methods also has been used to reduce models in mathematical epidemiology and other areas of mathematical biology in order to determine the long-term behavior of complicated models [29–33, 45–47]. For (2.9) and (2.13), however, we do not assume that there is a parametrizable time-scale separation between the original and emerging strain; in fact, such an assumption would be poorly justified for competing strains of COVID-19 as the infectivity and temporary immunity periods are on similar orders of magnitude. Consequently, we cannot treat the reduced system (2.13) as a asymptotically limiting case of (2.9). Nevertheless, numerical simulations show that the full system (2.9) and reduced system (2.13) have comparable dynamics when the full system (2.9) is assumed to start near the endemic steady state (see Figure 1). We leave further consideration of the relationship between the behaviors of the full model (2.9) and reduced model (2.13) as future work.

2.5. Reduced model analysis (all parameters)

We now turn our attention to the dynamical behavior of the switching system of differential equations (2.13). First, we want to ensure that the model is well-behaved and that solutions remain physically meaningful at all times. This is verified by the following result.

Theorem 2. *Solutions to the reduced two-strain model* (2.13) *starting in the following compact set* Λ *exist, are unique, and remain in* Λ *for all time* $t \geq 0$:

$$\Lambda = \{(I_2, R_2) \in \mathbb{R}^2_{\geq 0} \mid I_2 + R_2 \leq N\}. \tag{2.14}$$

Proof. We first show that Λ (2.14) is a trapping region of (2.13) with three boundaries: $I_2 = 0$, $R_2 = 0$, and $I_1 + R_2 = N$. Consider the boundary $I_2 = 0$. Substituting $I_2 = 0$ into (2.13) gives $I'_2 = 0$. It follows that the R_2-axis is an invariant set. Therefore, no solution may pass through $I_2 = 0$. Next, consider the boundary $R_2 = 0$. Substituting $R_2 = 0$ into (2.13) gives $R'_2 = \gamma_2 I_2 > 0$ for $I_2 > 0$. It follows that trajectories cannot escape the positive quadrant through $R_2 = 0$. Lastly, consider the boundary $I_2 + R_2 = N$ and the function $L(t) = I_2(t) + R_2(t)$. Along trajectories $(I_2(t), R_2(t))$ of (2.9) we have that

$$L'(t) = I'_2(t) + R'_2(t) = \frac{\beta_2}{N}(N - \omega(I_2, R_2) - I_2 - R_2)I_2 - \sigma_2 R_2 - \frac{\beta_1}{N}(1 - \epsilon)\omega(I_2, R_2)R_2$$

$$= -\frac{\beta_2}{N}\omega(I_2, R_2)I_2 - \sigma_2 R_2 - \frac{\beta_1}{N}(1 - \epsilon)\omega(I_2, R_2)R_2$$

where we have used the consideration that we are only interested in $I_2 + R_2 = N$ to reduce the first term. Since $\omega(I_2, R_2) \geq 0$ by (2.12), we have that $L'(t) < 0$ whenever $L(t) = N$. Consequently, trajectories starting in the set Λ given by (2.14) remain in Λ, and we are done.

To prove existence and uniqueness with Λ, we note that the vector field (2.13) is continuous in Λ and that Λ is closed and bounded and therefore a compact set. It follows that the vector field is Lipschitz continuous within Λ which implies that solutions exist and are unique, and we are done.

2.6. Reduced model analysis ($\epsilon = 0$ or $\epsilon = 1$)

We now consider the dynamics of (2.13) within the invariant set Λ in the specific cases of $\epsilon = 0$ (no cross-immunity) and $\epsilon = 1$ (full cross-immunity). Note that the system is planar so that it is sufficient to consider properties of the nullclines. We have the following result, which we prove in Appendix C.

Theorem 3. *Consider the reduced two-strain with asymmetric temporary immunity periods system (2.13) and basic reproduction numbers $\mathscr{R}_1$, $\mathscr{R}_2$, $\mathscr{R}_{12}$, and $\mathscr{R}_{21}$ as in (2.10). Suppose that $\mathscr{R}_1 > 1$, $\mathscr{R}_2 > 1$, and $\mathscr{R}_{21} > 1$ and either $\epsilon = 0$ or $\epsilon = 1$. Then the system has the following properties in* int(Λ):

1) **Vector field:** *The I_2-nullcline is continuous, has a strictly positive R_2-intercept, and is strictly decreasing, and the R_2-nullcline is continuous, intercepts the R_2-axis at $R_2 = 0$, and is strictly increasing. Furthermore, the regions bound by the I_2- and R_2-nullclines have the following directional properties:*

   (a) *Above the I_2- and R_2- nullclines, $I_2' < 0$ and $R_2' < 0$.*

   (b) *Above the I_2-nullcline and below the R_2-nullcline, $I_2' < 0$ and $R_2' > 0$.*

   (c) *Below the I_2-nullcline and above the R_2-nullcline, $I_2' > 0$ and $R_2' < 0$.*

   (d) *Below the I_2- and R_2- nullclines, $I_2' > 0$ and $R_2' > 0$.*

2) **Steady state existence and stability:** *The I_2- and R_2-nullclines have a unique intersection, corresponding to a unique positive co-existence steady state $(\bar{I}_2, \bar{R}_2) \in$ int(Λ). This steady state is locally exponentially stable and the global attractor for trajectories in* int(Λ). *Furthermore, we have the following:*

   (a) *If $\mathscr{R}_{12} > 1$ then $0 < \bar{I}_2 + \epsilon \bar{R}_2 < N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$ and $\bar{I}_1 = \omega(\bar{I}_2, \bar{R}_2) > 0$.*

   (b) *If $\mathscr{R}_{12} < 1$ then $\bar{I}_2 + \epsilon \bar{R}_2 \geq N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$ and $\bar{I}_1 = \omega(\bar{I}_2, \bar{R}_2) = 0$.*

Theorem 3 establishes the uniqueness and global stability of a strictly positive steady state when $\mathscr{R}_1 > 1$, $\mathscr{R}_2 > 1$, and $\mathscr{R}_{21} > 1$. It follows that, in these conditions, the emerging strain always has the capacity to infiltrate and survive in a population. Theorem 3 furthermore defines where that steady state lies in the state space. The original strain is able to survive as the emerging strain is infiltrating the population so long as $\mathscr{R}_{12} > 1$ and will die off otherwise. Although Theorem 3 is only proven for the cases of $\epsilon = 0$ (no cross-immunity) and $\epsilon = 1$ (full cross-immunity), we conjecture that the conclusions hold for all values $0 \leq \epsilon \leq 1$.

3. Numerical results
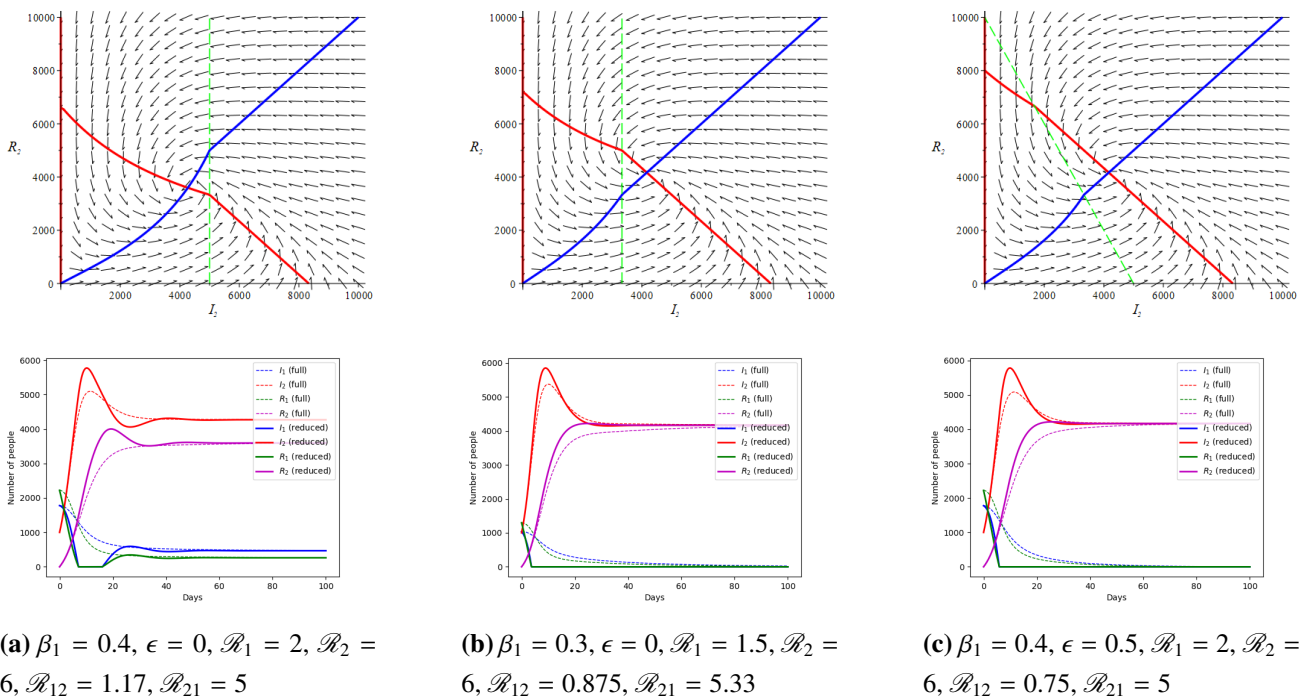
In this Section, we conduct some further analysis on the full and reduced two-strain with asymmetric temporary immunity period and partial cross immunity models (2.9) and (2.13).

3.1. Vector field plots and numerical simulations

We demonstrate the dynamical behavior of the full model (2.8) and reduced model (2.13) for three distinct sets of endemic parameters values in Figure 1.

The planar dynamics of the reduced system (2.13) may be represented with a vector field plot (Figure 1, upper row). We indicate the I_2-nullcline (red), R_2-nullcline (blue) and switching line $I_2 + \epsilon R_2 = N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$ (dashed green). The endemic steady state corresponds to the intersection of the nullclines. When this intersection is to the left of the green line, we have coexistence ($I_1 > 0$ and $I_2 > 0$) while when it is to the right we have emerging strain dominant behavior ($I_1 = 0$ and $I_2 > 0$). In the lower row of Figure 1, we include numerical simulations of the full system (2.9) (dashed) and the reduced system (2.13) (solid) for the same three sets of parameter values. The initial conditions $I_1(0)$ and $R_2(0)$ for the full system are chosen to be at the endemic steady state value according to (2.11).

(a) $\beta_1 = 0.4$, $\epsilon = 0$, $\mathscr{R}_1 = 2$, $\mathscr{R}_2 = 6$, $\mathscr{R}_{12} = 1.17$, $\mathscr{R}_{21} = 5$

(b) $\beta_1 = 0.3$, $\epsilon = 0$, $\mathscr{R}_1 = 1.5$, $\mathscr{R}_2 = 6$, $\mathscr{R}_{12} = 0.875$, $\mathscr{R}_{21} = 5.33$

(c) $\beta_1 = 0.4$, $\epsilon = 0.5$, $\mathscr{R}_1 = 2$, $\mathscr{R}_2 = 6$, $\mathscr{R}_{12} = 0.75$, $\mathscr{R}_{21} = 5$
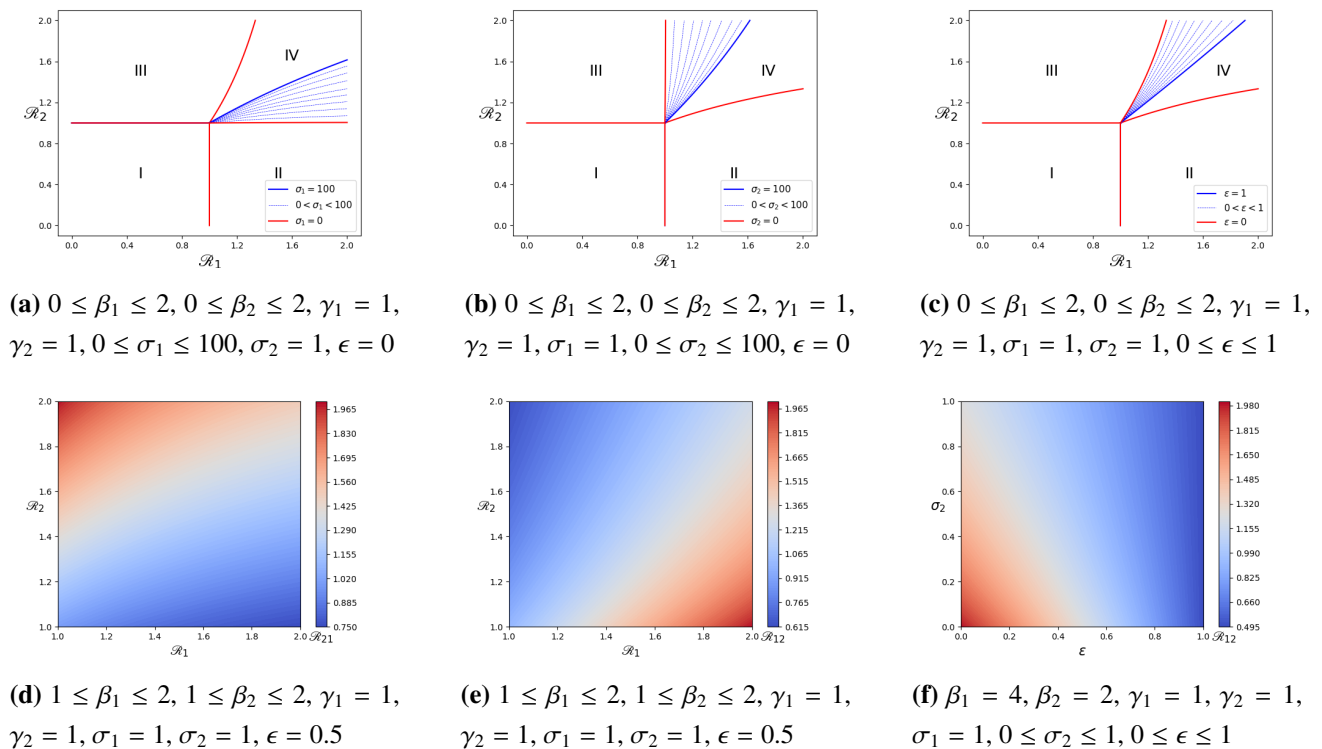
Figure 1. In the upper row, we show vector field plots of reduced model (2.13), and in the lower row, we show numerical simulations for the full (2.9) (dashed) and reduced (2.13) (solid) models. We utilize the parameter values $N = 10,000$, $\beta_2 = 0.6$, $\gamma_1 = 0.2$, $\gamma_2 = 0.1$, $\sigma_1 = 0.1$, $\sigma_2 = 0.1$, $I_2(0) = 1000$, and $R_2(0) = 0$. The remaining parameter values are indicated above. For the vector field plots, the I_2-nullcline (red), R_2-nullcline (blue), and switching line $I_2 + \epsilon R_2 = N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$ (dashed green) are indicated. The left of the transition line is the coexistence region ($I_1 > 0$ and $I_2 > 0$) while the right is the competitive exclusion region ($I_1 = 0$ and $I_2 > 0$). Notice that the number of people infected by the original strain ($I_1(t)$) in the reduced system (2.13) may hit zero and then become positive. This occurs in the vector field diagram when a trajectory transitions from the left of the switching line (green) to the right and then back to the left.

3.2. Bifurcation analysis

In Figure 2, we display selected bifurcation diagrams for the full model (2.9) and reduced model (2.13). In the upper row, we identify four qualitatively distinct regions of behavior in parameter space:

Region I—disease-free behavior where neither strain can infiltrate the population; Region II—original strain only behavior where the original strain may infiltrate the population but the emerging strain may not; Region III—emerging strain only behavior where the emerging strain may infiltrate the population but the original strain may not; Region IV—co-existence behavior where both strains may infiltrate the population. In the lower row, we represent how the multi-strain reproductive numbers $\mathscr{R}_{12}$ and $\mathscr{R}_{21}$ change as functions of other selected parameters.

(a) $0 \leq \beta_1 \leq 2, 0 \leq \beta_2 \leq 2, \gamma_1 = 1,$ $\gamma_2 = 1, 0 \leq \sigma_1 \leq 100, \sigma_2 = 1, \epsilon = 0$

(b) $0 \leq \beta_1 \leq 2, 0 \leq \beta_2 \leq 2, \gamma_1 = 1,$ $\gamma_2 = 1, \sigma_1 = 1, 0 \leq \sigma_2 \leq 100, \epsilon = 0$

(c) $0 \leq \beta_1 \leq 2, 0 \leq \beta_2 \leq 2, \gamma_1 = 1,$ $\gamma_2 = 1, \sigma_1 = 1, \sigma_2 = 1, 0 \leq \epsilon \leq 1$

(d) $1 \leq \beta_1 \leq 2, 1 \leq \beta_2 \leq 2, \gamma_1 = 1,$ $\gamma_2 = 1, \sigma_1 = 1, \sigma_2 = 1, \epsilon = 0.5$

(e) $1 \leq \beta_1 \leq 2, 1 \leq \beta_2 \leq 2, \gamma_1 = 1,$ $\gamma_2 = 1, \sigma_1 = 1, \sigma_2 = 1, \epsilon = 0.5$

(f) $\beta_1 = 4, \beta_2 = 2, \gamma_1 = 1, \gamma_2 = 1,$ $\sigma_1 = 1, 0 \leq \sigma_2 \leq 1, 0 \leq \epsilon \leq 1$
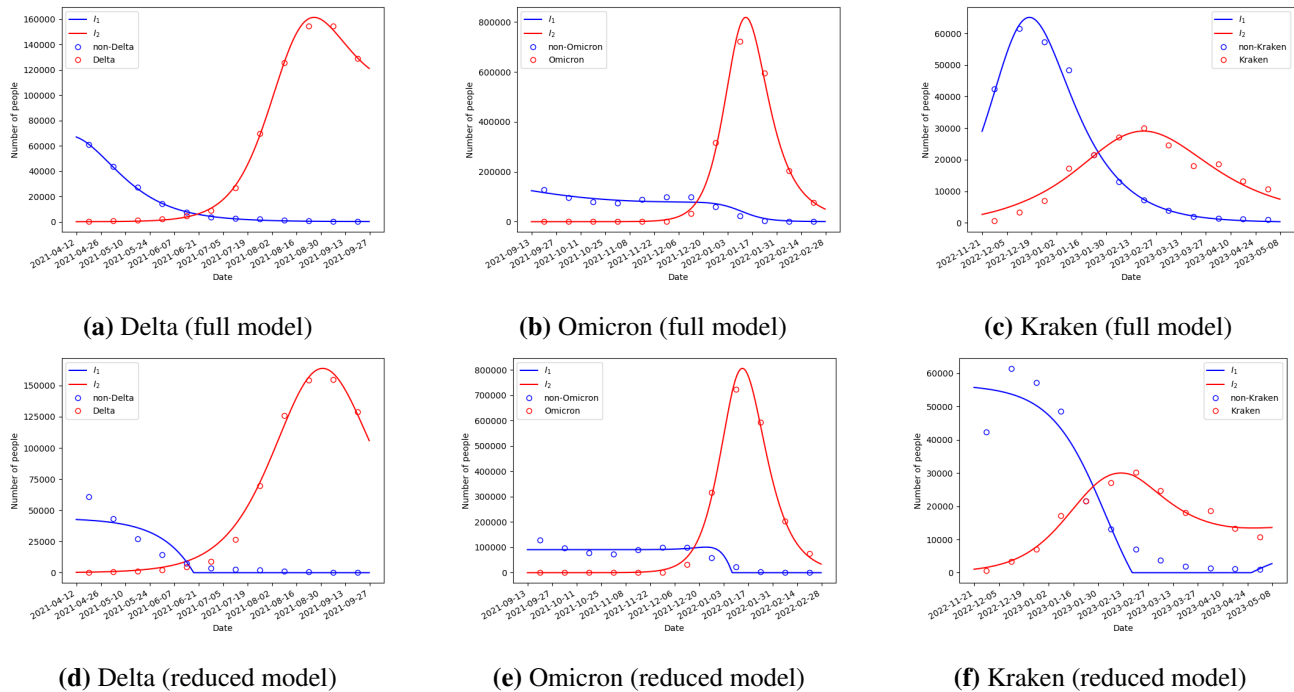
Figure 2. Bifurcation diagrams (upper row) and heat maps (lower row) for the steady states of (2.13). In the upper row, the parameter space is divided into four regions corresponding to four qualitatively distinct behaviors: Region I—disease-free behavior where neither strain can infiltrate the population; Region II—original strain only behavior where the original strain may infiltrate the population but the emerging strain may not; Region III—emerging strain only behavior where the emerging strain may infiltrate the population but the original strain may not; Region IV—co-existence behavior where both strains may infiltrate the population. The blue lines indicate how the regions change given changes in the indicated parameter. In the lower row, we display heat maps for the value of $\mathscr{R}_{21}$ (left) and $\mathscr{R}_{12}$ (center and right) as a function of various parameters. Note that the threshold $\mathscr{R}_{21} > 1$ is required for the emerging strain to infiltrate the population and $\mathscr{R}_{12} > 1$ is required for the original strain to survive if the emerging strain infiltrates. The parameters values and ranges are shown above.

3.3. Data fitting methods

To validate the differential equation models (2.9) and (2.13), we fit them to COVID-19 incidence and variant proportion data from the United States. COVID-19 case incidence data prior to 10/10/2022

was retrieved from the Johns Hopkins GitHub repository on May 5, 2023 [48] and after 10/10/2022 was retrieved from Our World in Data on June 5, 2023 [49]. Variant proportions data was retrieved from the Global Initiative on Sharing All Influenza Data (GISAID) on June 5, 2023 [50].

(a) Delta (full model)   **(b)** Omicron (full model)   **(c)** Kraken (full model)

(d) Delta (reduced model)   **(e)** Omicron (reduced model)   **(f)** Kraken (reduced model)

Figure 3. Comparisons of COVID-19 case incidence data to the best-fitting full model (2.9) and reduced model (2.13). In all figures, I_2 denotes the number of new cases of the variant of interest aggregated over the preceding two week period [**(a)** & **(d)**: Delta (B.1.617.2); **(b)** & **(e)**: Omicron (B.1.1.529); **(c)** & **(f)**: Omicron-subvariant Kraken (XBB.1.5)] and I_1 denotes the number of new cases over all other COVID-19 strains. Parameters are fit using the sum of squared error formula (3.1) in Python. COVID-19 case data is taken from the John Hopkins [48] and Our World in Data [49] and the variant data is taken from GISAID [50]. Data points are displayed at the midpoint of the two-week period during which they are aggregated. Best fitting parameters values can be found in Table 2.

We consider three separate periods of the COVID-19 pandemic during which they were a shift from the dominance of one variant to another in the United States: 1) the Delta takeover period (04/07/2021–08/16/2021); 2) the Omicron takeover period (08/30/2021–02/14/2022); and 3) the Omicron subvariant Kraken takeover period (11/21/2022–05/08/2023). In all cases, we consider the variant that is taking over as the emerging strain, I_2, and aggregate all remaining strains as the original strain, I_1. To determine variant numbers, we take the raw COVID-19 incidence data aggregated over the preceding two-week period and divide it according to the proportion of cases that are the emerging strain and which are the original strain. For optimizing the fit of the data to the model, we use the sum of squared error function

$$E(\mathbf{x}, \mathbf{y}; \theta) = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{3.1}$$

The values (x_i, y_i) are the cumulative clinically-confirmed COVID-19 cases over the preceding two

week period for the original (x) and emerging (y) strain, respectively, in the ith period of data collection. The values ($\hat{x}_i, \hat{y}_i$) are biweekly increase in new cases over the same time periods derived from the model. To simulate the model trajectories, we use the 4th order Runge-Kutta method. To fit the models to data, we use a customized stochastic Newton's method algorithm written in Python. We note that other COVID-19 studies have fit to variant proportions rather than estimated case numbers [20, 51].

Table 2. Best-fitting parameter values for the model simulations shown in Figure 3. For the full models (2.9) **(a)**–**(c)**, we fit the parameters N, $I_1(0)$, $R_1(0)$, $I_2(0)$, $R_2(0)$, β_1, β_2, γ_1, γ_2, σ_1, σ_2, and ϵ. For the reduced model (2.13) **(d)**–**(f)**, we fit the parameters N, $I_2(0)$, $R_2(0)$, β_1, β_2, γ_1, γ_2, σ_1, σ_2, and ϵ and determine the values of $I_1(0)$ and $R_1(0)$ from the system of Eqs (2.11). In order to focus on the potential impact of the temporary immunity periods σ_1^{-1} and σ_2^{-1} and the degree of cross-immunity ϵ, we imposed the restrictions $0.8\beta_1 \leq \beta_2 \leq 1.25\beta_1$, $\gamma_1 = \gamma_2$, $\sigma_1 \leq 0.5$, and $\sigma_2 \leq 0.5$. We also impose that all parameters are greater than or equal to 0.001.

Parameters	Delta		Omicron		Kraken	
	full **(a)**	reduced **(d)**	full **(b)**	reduced **(e)**	full **(c)**	reduced **(f)**
N	16,336,307	48,040,094	31,579,543	45,857,562	7,575,545	2,993,020
$I_1(0)$	188,363	188,363	1,462,667	813,673	96,790	283,535
$R_1(0)$	308,476	308,476	15,725,210	12	0	100
$I_2(0)$	85	722	1	2	9388	3661
$R_2(0)$	85	0	0	0	0	3661
β_1	0.36685	0.36316	0.18669	0.29794	0.30167	0.31235
β_2	0.43498	0.41680	0.23336	0.37243	0.28631	0.30993
γ_1	0.35878	0.36162	0.09294	0.24258	0.23641	0.22252
γ_2	0.35878	0.36162	0.09294	0.24258	0.23641	0.22252
σ_1	0.00266	0.5	0.00676	0.5	0.01421	0.09208
σ_2	0.04600	0.001	0.001	0.001	0.00544	0.02175
ϵ	0.001	1	1	1	1	1
$\mathscr{R}_1$	1.02250	1.00426	2.00875	1.22823	1.27603	1.40369
$\mathscr{R}_2$	1.21237	1.15260	2.51093	1.53528	1.21107	1.39284
$\mathscr{R}_{12}$	1.00199	0.87130	0.80000	0.80000	1.05364	1.00779
$\mathscr{R}_{21}$	1.21217	1.14977	2.42534	1.52273	1.19621	1.27560

In order to highlight the effects of the asymmetric temporary immunity periods and partial cross-immunity, we restrict the variant-specific transmission rates β_i and recovery rates γ_i to the limited ranges $0.8\beta_1 \leq \beta_2 \leq 1.25\beta_1$ and $\gamma_1 = \gamma_2$. That is, we assume that the emerging strain is no more or less than 25% as transmissible as the original strain and that the two strains have the same recovery period. Otherwise, we restrict all parameters from going to zero and $\sigma_i < 0.5$. For the full model fits (2.9), we fit over the parameters N, $I_1(0)$, $R_1(0)$, $I_2(0)$, $R_2(0)$, β_1, β_2, γ_1, γ_2, σ_1, σ_2, and ϵ. For the reduced model fits (2.13), we fit over the parameters N, $I_1(0)$, $R_1(0)$, $I_2(0)$, $R_2(0)$, β_1, β_2, γ_1, γ_2, σ_1, σ_2, and ϵ. Note that we also fit the population size N. This may seem unusual since the overall population size of the United States is well-known. The choice of fitting N was made to account for several factors which are not considered in the model: 1) removal of individuals from the susceptible population due to natural

immunity, quarantine, and vaccination; and 2) the incompleteness of COVID-19 case incidence data due to reporting errors and the prevalence of at-home self-test kits. The value N can be more fairly interpreted as the effective population level after accounting for the aggregate effects of removal from susceptibility and underreporting.

The best fitting model simulations are shown in Figure 3 and the corresponding best fitting parameters can be found in Table 2. The data circle corresponds to the average clinically-confirmed daily incidence of COVID-19 over the surrounding two week period. For ease of visualization, the data circle is displayed at the midpoint of the two-week data collection period.

4. Discussion

In this section, we analyze the results of the numerical studies carried out in Section 3.

4.1. Full and reduced model comparison

Consider the numerical simulations of the full model (2.9) and reduced model (2.13) (lower row of Figure 1). Note that there is no significant time-scale separation between the dynamics of the original and emerging strain so that the assumptions underlying the QSSA are not met. Consequently, we should not expect that trajectories of (2.9) converge to those of (2.13) in any limiting way. Nevertheless, our numerical study suggests that trajectories of the full and reduced systems agree well when the full system is taken to start near the endemic steady state of the original strain. After a short transient period, trajectories of both systems settle asymptotically to the same steady state. Since the long-term dynamics agree, we conjecture that the global stability of the steady states guaranteed by Theorem 3 holds not only for the reduced system (2.13) but also for the full system (2.9).

We now consider the performance of the full (2.9) and reduced (2.13) models when fit with COVID-19 case incidence data (see Figure 3).

1) During the Delta takeover period (Figure 3**(a)**,**(d)**), the full model fits the declining trend of the original strain (I_1) noticeably better than the reduced model. We also note a disparity in the survival prediction for the original strains ($\mathscr{R}_{12} > 1$ for full model and $\mathscr{R}_{12} < 1$ for the reduced model). This disparity can be attributed to a violation of the assumption underlying the derivation of the reduced model, which was that the original strain has reached its endemic steady state before the infiltration of the emerging strain.

2) During the Omicron takeover period (Figure 3**(b)**,**(e)**), the two models fit comparably well. In particular, the two models consistently predict that Omicron was more transmissible than previous strains ($\mathscr{R}_2 > \mathscr{R}_1$), that Omicron would infiltrate the population ($\mathscr{R}_{21} > 1$), and that the previous strains would not survive ($\mathscr{R}_{12} < 1$).

3) During the Kraken takeover period (Figure 3**(c)**,**(f)**), the two models fit the emerging strain (I_2) comparably well, while the reduced model fits the pre-takeover phase of the original strain (I_1) better and the full model fits the post-takeover phase of the original strain better. In particular, the reduced model better captures the steadiness of the original strain pre-takeover but overshoots its decline as the emerging strain enters the population. Both models, however, consistently predict the infiltration of the Kraken ($\mathscr{R}_{21} > 1$) and survival of the original strain ($\mathscr{R}_{12} > 1$).

Overall, the reduced model (2.13) performs well when fit with COVID-19 case incidence data. In particular, it fits the dynamics of the emerging strain (I_2) comparably well to the full mode. Care should be taken, however, to ensure the model assumption that the prevalence of existing strains is steady prior to the emergence of the new strain is met.

4.2. Effect of temporary immunity and partial cross-immunity

Now consider the bifurcation plots in Figure 2. Plots **(b)** and **(c)** show how changes in the loss of temporary immunity rates, σ_1 and σ_2, influence the boundaries of the four qualitatively distinct regions of parameter space. Note that as the temporary immunity periods become shorter (i.e., $\sigma_i \to \infty$, $\sigma_i^{-1} \to 0$), the co-existence region becomes smaller and the region for one-strain dominant behavior grows. This decrease in the capacity of coexistence as the temporary immunity periods decrease can be intuitively attributed to the increased amount of time where the two strains are competing for infections from a common susceptible class (S) rather than asymmetric pools of susceptibles ($S + R_1$ or $S + R_2$). In this case, it is more likely that the strictly more contagious strain will survive ($\mathscr{R}_1 > \mathscr{R}_2$ or $\mathscr{R}_2 > \mathscr{R}_1$). This analysis suggests that asymmetric temporary immunity periods is an important factor in the long-term survival of individual strains of COVID-19. Specifically, it is important for individual strains to be able to infect those recently infected with other strains to gain an advantage over those strains.

Plot **(c)** shows how changes in the degree of cross-immunity from the emerging strain to the original strain, ϵ, influences the boundaries of the regions. We can see that as the degree of cross-immunity increases ($\epsilon \to 1$), the region for coexistence shrinks dramatically. In this case, the original strain needs to have a strictly higher basic reproduction number in order to survive ($\mathscr{R}_1 > \mathscr{R}_2$). This can be intuitively interpreted as resulting from that reduction in individuals susceptible to the original strain (S) relative to the emerging strain ($S + R_1$). This suggests that even partial cross-immunity from the emerging strain to the original strain can contribute a significant competitive advantage to the emerging strain. Notably, an emerging strain of COVID-19 need not be significantly more contagious than an original strain (i.e., $\mathscr{R}_2 \gg \mathscr{R}_1$) in order for the new strain to eliminate the original strain from the population.

Plot **(d)** displays how the basic reproduction number for the emerging strain in the presence of the original strain $\mathscr{R}_{21}$ changes as a function of the strain-specific basic reproduction numbers $\mathscr{R}_1$ and $\mathscr{R}_2$. Having a more contagious emerging strain ($\mathscr{R}_2$ high) and a less contagious original strain ($\mathscr{R}_1$ low) makes it more likely for the emerging strain to be able to infiltrate the population ($\mathscr{R}_{21} > 1$). Plot **(e)** and **(f)** display how the basic reproduction number for the original strain in the presence of the emerging strain $\mathscr{R}_{12}$ changes as a function of $\mathscr{R}_1$, $\mathscr{R}_2$, σ_2, and ϵ. A more contagious original strain ($\mathscr{R}_1$ high), less contagious emerging strain ($\mathscr{R}_2$ low), long temporary immunity period for the emerging strain (σ_2^{-1} high) or low degree of cross-immunity (ϵ low) makes it more likely for the original strain to survive when the emerging strain infiltrates the population ($\mathscr{R}_{12} > 1$). The analysis suggests that the ability of COVID-19 strains to exclusively infect those recently infected by emerging strains is crucial for their long-term survival.

4.3. Implications for the spread of COVID-19

Notice that the parameter-fit strain-specific basic reproduction numbers $\mathscr{R}_1$ and $\mathscr{R}_2$ are not significantly different in any of the models (see Table 2). In fact, for the Kraken, the models predict that

the emerging strain is slightly less transmissible than the original strain ($\mathscr{R}_1 > \mathscr{R}_2$). This suggests that factors other than differences in transmissibility can be a driving force for the takeover of an emerging strain. For cases **(b)**–**(f)** in Figure 3 we have that the best fitting parameters include $\epsilon = 1$ while for case **(a)** we have $\sigma_1^{-1} > \sigma_2^{-1}$. This suggests that the emerging strain can gain a competitive advantage by increasing its cross-immunity or shortening its temporary immunity period. In these cases, even though the two strains have comparable strain-specific reproduction numbers, the emerging strain is able to infiltrate the already-infected population and drive the original strain to extinction or near-extinction.

For both the full and the reduced models, the data fitting suggests that the highest capacity of strain-over-strain fitness is for Omicron over previous strains ($\mathscr{R}_{21} = 2.42534$ for full model and $\mathscr{R}_{21} = 1.52273$ for reduced model). This is consistent with the data trend which showed Omicron accounting for over 99% of COVID-19 cases in the United States within three months of first being detected [50]. The data fitting also suggests that coexistence is possible with the Kraken strain and previous strains ($\mathscr{R}_{12} = 1.03467 > 1$ for full model and $\mathscr{R}_{21} = 1.10030 > 1$ for reduced model). Even though we have seen the prevalence of Kraken increase dramatically over the indicated time period, the model fits suggests the ancestral strains will survive rather than die off like previous takeovers. This is consistent with the observation that the Kraken variant has comprised approximate 90% of COVID-19 cases during post-takeover period of 03/27/2023–06/05/2023 [50]. This suggests that we are closer to having the long-term reality of co-circulating COVID-19 strains. Such a situation requires an adjustment of public health strategies to predict the most effective strain-specific vaccines, as is currently conducted with seasonal influenza.

We note that, while the basic reproduction numbers produced from fitting the models to data were consistent across optimization runs and similar for the full and reduced model, the individual parameters often varied significantly (see Table 2). For example, for the Delta strain, the full model fit produces $\epsilon \approx 0$ and $\sigma_1^{-1} \gg \sigma_2^{-1}$ while the reduced model fit produces $\epsilon = 1$ and $\sigma_2^{-1} \gg \sigma_1^{-1}$. Consequently, the full model parameter set suggests the variant takeover is due to a long temporary immunity period for the original strain, while the reduced model parameter set suggests it is due to complete cross-immunity of the emerging strain over the original strain. We hypothesize that this difference in plausible explanations of observed data is due to two primary factors: (a) the prevalence of many local minima of the highly nonlinear error function (3.1) over parameter space, many of which could be close to the global minimum; and (b) a lack of structural or practical parameter identifiability. The challenges of parameter identifiability for models of infectious disease spread have been increasingly acknowledged in the mathematical literature and are an ongoing area of study [52–54].

5. Conclusions and future work

We have introduced a two-strain model for infectious disease spread which incorporates asymmetric temporary immunity periods and partial cross-immunity. We have derived conditions (Theorem 2) on the temporary immunity periods, degree of cross-immunity, and basic reproduction numbers under which the strains can coexist. We have also reduced the model to a planar hybrid switching system and analyzed the dynamics using linear stability analysis, phase plane analysis, and the Bendixson-Dulac criterion. We have parameter fit the full model (2.9) and reduced model (2.13) to COVID-19 case incidence data from the United States for three different strains. This analysis has demonstrated the capacity of differences in temporary immunity periods and partial cross-immunity to account for the

observed changes in variant proportions over time, and in particular the phenomenon of one variant infiltrating a population and eliminating previous variants.

The work conducted in this paper suggests several fruitful opportunities for future work:

1) *Incorporating vaccination, births and deaths, and further strains.* Data has suggested, in particular, that vaccination has played a significant role in altering the spread of COVID-19 [11]. Future work will incorporate vaccination and other factors in the models to the impact of asymmetries in vaccine-resistance between strains in promoting competitive exclusion or coexistence.

2) *Extending to a multiscale within-host and between-host framework.* A more accurate approach for incorporating host immunity and variances in cross-variant immunity is by considering within-host dynamics. Incorporating within-host dynamics with between-host population dynamics requires a multiscale approach, which is a growing area of research [55–57]. This will be a focus of future research.

3) *Extending dynamical results to the full model* (2.9) *and the reduced model* (2.13) *for* $0 \le \epsilon \le 1$. It is suspected that the conclusions of Theorem 3 hold for the reduced model (2.13) for all values of $0 \le \epsilon \le 1$, rather than just the special cases $\epsilon = 0$ and $\epsilon = 1$, and also the full model (2.9). These results, however, remain unproved. A sufficient condition for Theorem 3 to hold for the reduced model (2.13) would be

$$\frac{\partial \omega}{\partial R_2} R + \omega(I_2, R_2) \ge 0$$

for $0 < I_2 + \epsilon R_2 < N\left(1 - \frac{1}{\mathscr{R}_1}\right)$. Future work will aim to complete this analysis.

Acknowledgments

Conflicts of interest

The authors declare no conflict of interest.

References

1. I. A. Baba, A. Yusuf, K. S. Nisar, A. H. Abdel-Aty, T. A. Nofal, Mathematical model to assess the imposition of lockdown during COVID-19 pandemic, *Results Phys.*, **20** (2021), 103716. https://doi.org/10.1016/j.rinp.2020.103716

2. T. Sardar, S. K. Nadim, J. Chattopadhyay, Assessment of lockdown effect in some states and overall India: A predictive mathematical study on COVID-19 outbreak, *Chaos Solitons Fractals*, **139** (2020), 110078. https://doi.org/10.1016/j.chaos.2020.110078

3. N. Wang, Y. Fu, H. Zhang, H. Shi, An evaluation of mathematical models for the outbreak of COVID-19, *Precis. Clin. Med.*, **3** (2020), 85–93. https://doi.org/10.1093/pcmedi/pbaa016

4. J. T. Wu, S. Mei, S. Luo, K. Leung, D. Liu, Q. Lv, et al., A global assessment of the impact of school closure in reducing COVID-19 spread, *Philos. Trans. Royal Soc. A*, **380** (2022), 20210124. https://doi.org/10.1098/rsta.2021.0124

5. T. A. Biala, Y. O. Afolabi, A. Q. M. Khaliq, How efficient is contact tracing in mitigating the spread of COVID-19? A mathematical modeling approach, *Appl. Math. Model.*, **103** (2022), 714–730. https://doi.org/10.1016/j.apm.2021.11.011

6. A. Bilinski, F. Mostashari, J. A. Salomon, Modeling contact tracing strategies for COVID-19 in the context of relaxed physical distancing measures, *JAMA Netw. Open*, **3** (2020), e2019217–e2019217. https://doi.org/10.1001/jamanetworkopen.2020.19217

7. S. E. Eikenberry, M. Mancuso, E. Iboi, T. Phan, K. Eikenberry, Y. Kuang, et al., To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic, *Infect. Dis. Model.*, **5** (2020), 293–308. https://doi.org/10.1016/j.idm.2020.04.001

8. C. N. Ngonghala, E. Iboi, S. Eikenberry, M. Scotch, C. R. MacIntyre, M. H. Bonds, et al., Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel coronavirus, *Math. Biosci.*, **325** (2020), 108364. https://doi.org/10.1016/j.mbs.2020.108364

9. B. H. Foy, B. Wahl, K. Mehta, A. Shet, G. I. Menon, C. Britto, Comparing COVID-19 vaccine allocation strategies in India: A mathematical modelling study, *Int. J. Infect. Dis.*, **103** (2021), 431–438. https://doi.org/10.1016/j.ijid.2020.12.075

10. K. Liu, Y. Lou, Optimizing COVID-19 vaccination programs during vaccine shortages: A review of mathematical models, *Infect. Dis. Model.*, **7** (2022), 286–298. https://doi.org/10.1016/j.idm.2022.02.002

11. M. D. Johnston, B. Pell, P. Nelson, A mathematical study of COVID-19 spread by vaccination status in Virginia, *Appl. Sci.*, **12** (2022), 1723. https://doi.org/10.3390/app12031723

12. L. Lin, Y. Zhao, B. Chen, D. He, Multiple COVID-19 waves and vaccination effectiveness in the united states, *Int. J. Environ. Res. Public Health*, **19** (2022), 2282. https://doi.org/10.3390/ijerph19042282

13. B. Pell, S. Brozak, T. Phan, F. Wu, Y. Kuang, The emergence of a virus variant: dynamics of a competition model with cross-immunity time-delay validated by wastewater surveillance data for COVID-19, *J. Math. Biol.*, **86** (2023), 63. https://doi.org/10.1007/s00285-023-01900-0

14. M. Fudolig, R. Howard, The local stability of a modified multi-strain sir model for emerging viral strains, *PLoS One*, **15** (2020), e0243408. https://doi.org/10.1371/journal.pone.0243408

15. W. O. Kermack, A. G. McKendrick, A contribution to the mathematical theory of epidemics, *Proc. R. Soc. London, Ser. A*, **115** (1927), 700–721. https://doi.org/10.1016/S0092-8240(05)80040-0

16. V. Andreasen, J. Lin, S. A. Levin, The dynamics of cocirculating influenza strains conferring partial cross-immunity, *J. Math. Biol.*, **35** (1997), 825–842. https://doi.org/10.1007/s002850050079

17. H. W. Hethcote, The mathematics of infectious diseases, *SIAM Rev. Soc. Ind. Appl. Math.*, **42** (2000), 599–653. https://doi.org/10.1137/S0036144500371907

18. J. Lin, V. Andreasen, S. A. Levin, Dynamics of influenza a drift: the linear three-strain model, *Math. Biosci.*, **162** (1999), 33–51. https://doi.org/10.1016/S0025-5564(99)00042-5

19. L. J. White, M. J. Cox, G. F. Medley, Cross immunity and vaccination against multiple microparasite strains, *Math. Med. Biol.*, **15** (1998), 211–233. https://doi.org/10.1093/imammb/15.3.211

20. A. S. Ciupeanu, M. Varughese, W. C. Roda, D. Han, Q. Cheng, M. Y. Li, Mathematical modeling of the dynamics of COVID-19 variants of concern: Asymptotic and finite-time perspectives, *Infect. Dis. Model.*, **7** (2022), 581–596. https://doi.org/10.1016/j.idm.2022.08.004

21. W. Wang, Competitive exclusion of two viral strains of COVID-19, *Infect. Dis. Model.*, **7** (2022), 637–644. https://doi.org/10.1016/j.idm.2022.10.001

22. Y. Wang, J. Ma, J. Cao, Basic reproduction number for the SIR epidemic in degree correlated networks, *Physica D*, **433** (2022), 133183. https://doi.org/10.1016/j.physd.2022.133183

23. M. Martcheva, *An Introduction to Mathematical Epidemiology*, Springer, New York, **61** (2015), 9–31.

24. O. Khyar, K. Allali, Global dynamics of a multi-strain SEIR epidemic model with general incidence rates: application to COVID-19 pandemic, *Nonlinear Dyn.*, **102** (2020), 489–509. https://doi.org/10.1007/s11071-020-05929-4

25. X. Wang, J. Yang, X. Luo, Competitive exclusion and coexistence phenomena of a two-strain SIS model on complex networks from global perspectives, *J. Appl. Math. Comput.*, **68** (2022), 1–19. https://doi.org/10.1007/s12190-022-01712-3

26. B. P. Ingalls, *Mathematical Modeling in Systems Biology: An Introduction*, MIT press, 2013.

27. L. Michaelis, M. L. Menten, Die kinetik der invertinwirkung, *Biochem. Z*, **49** (1913), 333–369.

28. A. V. Hill, The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves, *J. Physiol.*, **40** (1910), iv–vii.

29. C. Castillo-Chavez, Z. Feng, To treat or not to treat: the case of tuberculosis, *J. Math. Biol.*, **35** (1997), 629–656. https://doi.org/10.1007/s002850050069

30. K. Wang, Y. Kuang, Novel dynamics of a simple daphnia-microparasite model with dose-dependent infection, *Discrete Contin. Dyn. Syst. Ser. A*, **4** (2011), 1599–1610. https://doi.org/10.3934/dcdss.2011.4.1599

31. E. M. Rutter, Y. Kuang, Global dynamics of a model of joint hormone treatment with dendritic cell vaccine for prostate cancer, *Discrete Contin. Dyn. Syst. Ser. A*, **22** (2017), 1001. https://doi.org/10.3934/dcdsb.2017050

32. P. Rashkov, E. Venturino, M. Aguiar, N. Stollenwerk, B. W. Kooi, On the role of vector modeling in a minimalistic epidemic model, *Math. Biosci. Eng.*, **16** (2019), 4314–4338. https://doi.org/10.3934/mbe.2019215

33. P. Rashkov, B. W. Kooi, Complexity of host-vector dynamics in a two-strain dengue model, *J. Biol. Dyn.*, **15** (2021), 35–72. https://doi.org/10.1080/17513758.2020.1864038

34. O. Diekmann, J. A. P. Heesterbeek, On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations, *J. Math. Biol.*, **28** (1990), 365–382.

35. P. van den Driessche, J. Watmough, Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission, *Math. Biosci.*, **180** (2002), 29–48. https://doi.org/10.1016/S0025-5564(02)00108-6

36. O. Diekmann, J. A. P. Heesterbeek, M. G. Roberts, The construction of next-generation matrices for compartmental epidemic models, *J. R. Soc. Interface*, **7** (2009), rsif20090386. https://doi.org/10.1098/rsif.2009.0386

37. J. M. Heffernan, R. J. Smith, L. M. Wahl, Perspectives on the basic reproductive ratio, *J. R. Soc. Interface*, **2** (2005), 281–293. https://doi.org/10.1098/rsif.2005.0042

38. P. van den Driessche, J. Watmough, Further notes on the basic reproduction number, in *Mathematical Epidemiology* (eds. F. Brauer, P. Driessche and J. Wu), (2008), 159–178. https://doi.org/10.1007/978-3-540-78911-6

39. P. van den Driessche, Reproduction numbers of infectious disease models, *Infect. Dis. Model.*, **2** (2017), 288–303. https://doi.org/10.1016/j.idm.2017.06.002

40. X. Liu, H. Lin, B. M. Chen, Structural controllability of switched linear systems, *Automatica*, **49** (2013), 3531–3537. https://doi.org/10.1016/j.automatica.2013.09.015

41. M. Egerstedt, Y. Wardi, F. Delmotte, Optimal control of switching times in switched dynamical systems, in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03ch37475)*, IEEE, **3** (2003), 2138–2143. https://doi.org/10.1109/CDC.2003.1272934

42. C. Sun, W. Yang, J. Arino, K. Khan, Effect of media-induced social distancing on disease transmission in a two patch setting, *Math. Biosci.*, **230** (2011), 87–95. https://doi.org/10.1016/j.mbs.2011.01.005

43. F. Zhu, P. J. Antsaklis, Optimal control of hybrid switched systems: A brief survey, *Discrete Event Dyn. Syst.*, **25** (2015), 345–364. https://doi.org/10.1007/s10626-014-0187-5

44. T. A. Burton, *Volterra Integral and Differential Equations*, Elsevier, 2005.

45. E. Feliu, C. Wiuf, Simplifying biochemical models with intermediate species, *J. R. Soc. Interface*, **10** (2013), 20130484. https://doi.org/10.1098/rsif.2013.0484

46. E. Feliu, D. Lax, S. Walcher, C. Wiuf, Quasi-Steady-State and singular perturbation reduction for reaction networks with noninteracting species, *SIAM J. Appl. Dyn. Syst.*, 21 (2022), 782–816. https://doi.org/10.1137/20M1364503

47. H. R. Thieme, Convergence results and a Poincaré-Bendixson trichotomy for asymptotically autonomous differential equations, *J. Math. Biol.*, **30** (1992), 755–763. https://doi.org/10.1007/BF00173267

48. L. Gardner, E. Dong, H. Du, An interactive web-based dashboard to track COVID-19 in real time, *Lancet Infect. Dis.*, **20** (2023), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

49. E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, et al., Coronavirus pandemic (COVID-19), *Our World Data*, 2020.

50. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data–from vision to reality, *Eurosurveillance*, **22** (2017), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

51. L. Boyle, S. Hletko, H. Huang, J. Lee, G. Pallod, H. R. Tung, et al., Selective sweeps in SARS-CoV-2 variant competition, *Proc. Natl. Acad. Sci.*, **119** (2002), e2213879119. https://doi.org/10.1073/pnas.2213879119

52. Y. H. Kao, M. C. Eisenberg, Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment, *Epidemics*, **25** (2018), 89–100. https://doi.org/10.1016/j.epidem.2018.05.010

53. G. Massonis, J. R. Banga, A. F. Villaverde, Structural identifiability and observability of compartmental models of the COVID-19 pandemic, *Annu. Rev. Control*, **51** (2021), 441–459. https://doi.org/10.1016/j.arcontrol.2020.12.001

54. N. Tuncer, A. Timsina, M. Nuno, G. Chowell, M. Martcheva, Parameter identifiability and optimal control of an SARS-CoV-2 model early in the pandemic, *J. Biol. Dyn.*, **16** (2022), 412–438. https://doi.org/10.1080/17513758.2022.2078899

55. M. O. Adewole, T. S. Faniran, F. A. Abdullah, M. K. Ali, COVID-19 dynamics and immune response: Linking within-host and between-host dynamics, *Chaos Solitons Fractals*, **2023** (2023), 113722. https://doi.org/10.1016/j.chaos.2023.113722

56. X. Wang, S. Wang, J. Wang, L. Rong, A multiscale model of COVID-19 dynamics, *Bull. Math. Biol.*, **84** (2022), 99. https://doi.org/10.1007/s11538-022-01058-8

57. Y. Xue, D. Chen, S. R. Smith, X. Ruan, S. Tang, Coupling the within-host process and between-host transmission of COVID-19 suggests vaccination and school closures are critical, *Bull. Math. Biol.*, **85** (2023), 6. https://doi.org/10.1007/s11538-023-01132-9

58. S. Wiggins, M. Golubitsky, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer, New Delhi, 2003.

59. D. M. Grobman, Homeomorphism of systems of differential equations, *Dokl. Akad. Nauk SSSR*, **128** (1959), 880–881.

60. P. Hartman, A lemma in the theory of structural stability of differential equations, *Proc. Am. Math. Soc.*, **11** (1960), 610–620. https://doi.org/10.1090/S0002-9939-1960-0121542-7

Appendix

A. Next generation method

Following [35], we define the state vector $\mathbf{x} \in \mathbb{R}_{\geq 0}^n$ and reindex so that $i = 1, \ldots, m$, $m \leq n$, corresponds to the infectious compartments. For $i = 1, \ldots, m$, let $\mathscr{F}_i(\mathbf{x})$ denote the rate of new infections in compartment i, $\mathscr{V}_i^+(\mathbf{x})$ denote the rate of inflow into compartment i by any other means, and $\mathscr{V}_i^-(\mathbf{x})$ denote the rate of outflow from compartment i by any other means. The rate of change in each infectious compartment $i = 1, \ldots, m$, can then be given by

$$\frac{dx_i}{dt} = \mathscr{F}_i - \mathscr{V}_i$$

where $\mathscr{V}_i = \mathscr{V}_i^- - \mathscr{V}_i^+$. Let

$$F = \left[\frac{d\mathscr{F}_i}{dx_j}\right], \ i, j = 1, \ldots, m$$

and

$$V = \left[\frac{d\mathcal{V}_i}{dx_j} \right], \ i, j = 1, \ldots, m$$

denote the Jacobians of $\mathcal{F}$ and $\mathcal{V}$, respectively, restricted to only infectious compartments. The next generation matrix is given by FV^{-1}. If $\mathbf{x}_0$ is the disease free steady state, the basic reproduction number is given by $\mathcal{R}_0 = \rho(FV^{-1}(\mathbf{x}_0))$. If FV^{-1} is reducible with n irreducible components, we can decouple the system into strain specific submatrices $F_i V_i^{-1}$. We can then compute the basic reproduction number of strain i by $\mathcal{R}_i = \rho(F_i V_i^{-1}(\mathbf{x}_0))$ and derive $\mathcal{R}_0 = \max_{i=1,\ldots,m} \{\mathcal{R}_i\}$. Similarly, if $\mathbf{x}_j$ is the strain j only steady state, then the basic reproduction number of strain i in the presence of strain j is given by $\mathcal{R}_{ij} = \rho(F_i V_i^{-1}(\mathbf{x}_j))$.

A.1. Basic two-strain model

For the basic two-strain model (2.1), we have the following:

$$\mathcal{F} = \left[\begin{array}{c} \frac{\beta_1}{N} S I_1 \\ \frac{\beta_2}{N} S I_2 \end{array} \right] \text{ and } \mathcal{V} = \left[\begin{array}{c} \gamma_1 I_1 \\ \gamma_2 I_2 \end{array} \right]$$

so that

$$F = \left[\begin{array}{cc} \frac{\beta_1}{N} S & 0 \\ 0 & \frac{\beta_2}{N} S \end{array} \right] \text{ and } V = \left[\begin{array}{cc} \gamma_1 & 0 \\ 0 & \gamma_2 \end{array} \right].$$

It follows that

$$FV^{-1} = \left[\begin{array}{cc} \frac{\beta_1}{\gamma_1 N} S & 0 \\ 0 & \frac{\beta_2}{\gamma_2 N} S \end{array} \right], F_1 V_1^{-1} = \frac{\beta_1}{\gamma_1 N} S, \text{ and } F_2 V_2^{-1} = \frac{\beta_2}{\gamma_2 N} S$$

where we have removed the matrix representation for the 1×1 matrices for notational simplicity. Substituting in the disease free steady state $\mathbf{x}_0$ (2.4) gives

$$FV^{-1}(\mathbf{x}_0) = \left[\begin{array}{cc} \frac{\beta_1}{\gamma_1} & 0 \\ 0 & \frac{\beta_2}{\gamma_2} \end{array} \right], F_1 V_1^{-1}(\mathbf{x}_0) = \frac{\beta_1}{\gamma_1}, \text{ and } F_2 V_2^{-1}(\mathbf{x}_0) = \frac{\beta_2}{\gamma_2}.$$

It follows that $\mathcal{R}_1 = \beta_1/\gamma_1$, $\mathcal{R}_2 = \beta_2/\gamma_2$, and $\mathcal{R}_0 = \max\{\mathcal{R}_1, \mathcal{R}_2\}$. Furthermore, substituting in the original strain only steady state $\mathbf{x}_1$ (2.5) and emerging strain only steady state $\mathbf{x}_2$ (2.6) into the relevant components gives

$$F_1 V_1^{-1}(\mathbf{x}_2) = \frac{\beta_1 \gamma_2}{\beta_2 \gamma_1}, \text{ and } F_2 V_2^{-1}(\mathbf{x}_1) = \frac{\beta_2 \gamma_1}{\beta_1 \gamma_2}.$$

It follows that $\mathcal{R}_{12} = \mathcal{R}_1/\mathcal{R}_2$ and $\mathcal{R}_{21} = \mathcal{R}_2/\mathcal{R}_1$ and we are done.

A.2. Two-strain model with asymmetric temporary immunity periods and partial cross-immunity

For the two-strain model with asymmetric temporary immunity periods and partial cross-immunity (2.7), we have the following:

$$\mathscr{F} = \begin{bmatrix} \dfrac{\beta_1}{N}(S + (1-\epsilon)R_2)I_1 \\[2ex] \dfrac{\beta_2}{N}(S + R_1)I_2 \end{bmatrix} \text{ and } \mathscr{V} = \begin{bmatrix} \gamma_1 I_1 \\[1ex] \gamma_2 I_2 \end{bmatrix}$$

so that

$$F = \begin{bmatrix} \dfrac{\beta_1}{N}(S + (1-\epsilon)R_2) & 0 \\[2ex] 0 & \dfrac{\beta_2}{N}(S + R_1) \end{bmatrix} \text{ and } V = \begin{bmatrix} \gamma_1 & 0 \\[1ex] 0 & \gamma_2 \end{bmatrix}.$$

It follows that

$$FV^{-1} = \begin{bmatrix} \dfrac{\beta_1}{\gamma_1 N}(S + (1-\epsilon)R_2) & 0 \\[2ex] 0 & \dfrac{\beta_2}{\gamma_2 N}(S + R_1) \end{bmatrix}, F_1 V_1^{-1} = \dfrac{\beta_1}{\gamma_1 N}(S + (1-\epsilon)R_2), \text{ and } F_2 V_2^{-1} = \dfrac{\beta_2}{\gamma_2 N}(S + R_1).$$

Substituting in the disease free steady state $\mathbf{x}_0$ (2.4) gives

$$FV^{-1}(\mathbf{x}_0) = \begin{bmatrix} \dfrac{\beta_1}{\gamma_1} & 0 \\[2ex] 0 & \dfrac{\beta_2}{\gamma_2} \end{bmatrix}, F_1 V_1^{-1}(\mathbf{x}_0) = \dfrac{\beta_1}{\gamma_1}, \text{ and } F_2 V_2^{-1}(\mathbf{x}_0) = \dfrac{\beta_2}{\gamma_2}.$$

It follows that $\mathscr{R}_1 = \beta_1/\gamma_1$, $\mathscr{R}_2 = \beta_2/\gamma_2$, and $\mathscr{R}_0 = \max\{\mathscr{R}_1, \mathscr{R}_2\}$. Now, substituting in the original strain only steady state $\mathbf{x}_1$ (2.5) and emerging strain only steady state $\mathbf{x}_2$ (2.6) into the relevant components gives

$$F_1 V_1^{-1}(\mathbf{x}_2) = \frac{\beta_1\gamma_2((1-\epsilon)\beta_2 + \epsilon\gamma_2 + \sigma_2)}{\beta_2\gamma_1(\gamma_2 + \sigma_2)}, \text{ and } F_2 V_2^{-1}(\mathbf{x}_1) = \frac{\beta_2\gamma_1(\beta_1 + \sigma_1)}{\beta_1\gamma_2(\gamma_1 + \sigma_1)}.$$

It follows that $\mathscr{R}_{12} = \dfrac{\mathscr{R}_1}{\mathscr{R}_2}\left(\dfrac{(1-\epsilon)\beta_2 + \epsilon\gamma_2 + \sigma_2}{\gamma_2 + \sigma_2}\right)$ and $\mathscr{R}_{21} = \dfrac{\mathscr{R}_2}{\mathscr{R}_1}\left(\dfrac{\beta_1 + \sigma_1}{\gamma_1 + \sigma_1}\right)$ and we are done.

B. Proof of Theorem 1

Proof. We want to prove that a co-existence steady state (i.e., $I_1 > 0$ and $I_2 > 0$) exists if and only if $\min\{\mathscr{R}_1, \mathscr{R}_2, \mathscr{R}_{12}, \mathscr{R}_{21}\} > 1$ is satisfied.

We consider the steady state equations for (2.9):

$$\begin{cases} \dfrac{\beta_1}{N}(N - I_1 - R_1 - I_2 - \epsilon R_2)I_1 - \gamma_1 I_1 = 0 \\[2ex] \gamma_1 I_1 - \sigma_1 R_1 - \dfrac{\beta_2}{N}R_1 I_2 = 0 \\[2ex] \dfrac{\beta_2}{N}(N - I_1 - I_2 - R_2)I_2 - \gamma_2 I_2 = 0 \\[2ex] \gamma_2 I_2 - \sigma_2 R_2 - \dfrac{\beta_1}{N}(1-\epsilon)R_2 I_1 = 0 \end{cases} \quad \text{(B.1)}$$

We can solve the latter two equations in (B.1) to get

$$I_2 = \frac{(N(\beta_2 - \gamma_2) - \beta_2 I_1)(\beta_1(1 - \epsilon)I_1 + N\sigma_2)}{\beta_2(\beta_1(1 - \epsilon)I_1 + N(\gamma_2 + \sigma_2))} =: \phi(I_1),$$

$$R_2 = \frac{N\gamma_2(N(\beta_2 - \gamma_2) - \beta_2 I_1)}{\beta_2(\beta_1(1 - \epsilon)I_1 + N(\gamma_2 + \sigma_2))} =: \psi(I_1) \tag{B.2}$$

We note that this steady state only satisfies $I_1 > 0$ and $R_1 > 0$ if $I_1 < N\left(1 - \dfrac{1}{\mathscr{R}_2}\right)$.

We now substitute the equations $\phi(I_1)$ and $\psi(I_1)$ from (B.2) into the first two equations of (B.1). This gives

$$\begin{cases} \dfrac{\beta_1}{N}(N - I_1 - \phi(I_1) - R_1 - \epsilon\psi(I_1))I_1 - \gamma_1 I_1 = 0 \\[2mm] \gamma_1 I_1 - \sigma_1 R_1 - \dfrac{\beta_2}{N}R_1\phi(I_1) = 0. \end{cases} \tag{B.3}$$

In order to show whether these two equations can be satisfied for $I_1 > 0$, we solve for R_1 in (B.3) in terms of I_1 and compute the corresponding derivatives. We have the following:

$$\begin{cases} R_1 = N - I_1 - \phi(I_1) - \epsilon\psi(I_1) - \dfrac{N\gamma_1}{\beta_1} =: f(I_1) \\[3mm] R_1 = \dfrac{N\gamma_1 I_1}{N\sigma_1 + \beta_2\phi(I_1)} =: g(I_1) \end{cases} \tag{B.4}$$

Computing and simplifying the derivatives gives the following:

$$\begin{cases} f'(I_1) = -\dfrac{N^2\gamma_2(1 - \epsilon)((1 - \epsilon)(\beta_2 - \gamma_2)\beta_1 + \beta_2(\gamma_2 + \sigma_2))}{\beta_2(\beta_1 I_1(1 - \epsilon) + N(\gamma_2 + \sigma_2))^2} \\[3mm] g'(I_1) = \dfrac{N\gamma_1([\beta_2(I_1^2(1 - \epsilon)^2\beta_1^2 + 2NI_1\sigma_2(1 - \epsilon)\beta_1 + N^2\sigma_2(\gamma_2 + \sigma_2))]\phi(I_1) + (\beta_1 I_1(1 - \epsilon) + N\sigma_2)(\beta_2\beta_1(1 - \epsilon)I_1^2 + (\sigma_1(1 - \epsilon)\beta_1 + \beta_2\sigma_2)NI_1 + N^2\sigma_1(\gamma_2 + \sigma_2)))}{(N\sigma_1 + \beta_2\phi(I_1))^2}. \end{cases}$$

Since $1 - \epsilon \geq 0, \beta_2 > \gamma_2$, and $\phi(I_1) \geq 0$, we have that $f'(I_1) < 0$ and $g'(I_1) > 0$.

We observe that $g(0) = 0$. In order to have a solution in the region $I_1 > 0$ and $R_1 > 0$, it is necessary that $f(0) > 0$, which corresponds to

$$\mathscr{R}_1 > \mathscr{R}_2\left(\frac{\gamma_2 + \sigma_2}{(1 - \epsilon)\beta_2 + \epsilon\gamma_2 + \sigma_2}\right) \implies \frac{\mathscr{R}_1}{\mathscr{R}_2}\left(\frac{(1 - \epsilon)\beta_2 + \epsilon\gamma_2 + \sigma_2}{\gamma_2 + \sigma_2}\right) > 1.$$

We now check where that solution can be. We will have a solution in the region $0 < I_1 < N\left(1 - \dfrac{1}{\mathscr{R}_2}\right) =: I_1^*$, corresponding to $I_2 > 0$, if and only if $g(I_1^*) > f(I_1^*)$. Substituting $I_1 = I_1^*$ into (B.4) gives the condition

$$\mathscr{R}_2 > \mathscr{R}_1\left(\frac{\gamma_1 + \sigma_1}{\beta_1 + \sigma_1}\right) \implies \frac{\mathscr{R}_2}{\mathscr{R}_1}\left(\frac{\beta_1 + \sigma_1}{\gamma_1 + \sigma_1}\right) > 1.$$

It follows that a solution to (B.3) with $I_2 > 0$ and $I_1 > 0$ if and only if $\min\{\mathscr{R}_1, \mathscr{R}_2, \mathscr{R}_{12}, \mathscr{R}_{21}\} > 1$ is satisfied.

C. Proof of Theorem 3

We start our analysis of (2.13) by deriving a few properties of the original strain steady state function $\omega(I_2, R_2)$ (2.12). We have the following Lemma.

Lemma 1. *The original strain steady state function $\omega(I_2, R_2)$ (2.12) satisfies the following properties:*

1)
$$\frac{\partial \omega}{\partial I_2} = \begin{cases} N\gamma_1 \left(\dfrac{1 + \dfrac{\beta_2 \omega(I_2, R_2)}{\beta_2 I_2 + N\sigma_1}}{\beta_2 I_2 + N(\gamma_1 + \sigma_1)} \right) - 1, & \text{if } I_2 + \epsilon R_2 < N\left(1 - \dfrac{1}{\mathscr{R}_1}\right) \\[4ex] 0, & \text{if } I_2 + \epsilon R_2 > N\left(1 - \dfrac{1}{\mathscr{R}_1}\right) \end{cases}$$

2)
$$\frac{\partial \omega}{\partial R_2} = \begin{cases} -\epsilon \left(\dfrac{\beta_2 I_2 + N\sigma_1}{\beta_2 I_2 + N(\gamma_1 + \sigma_1)} \right), & \text{if } I_2 + \epsilon R_2 < N\left(1 - \dfrac{1}{\mathscr{R}_1}\right) \\[3ex] 0, & \text{if } I_2 + \epsilon R_2 > N\left(1 - \dfrac{1}{\mathscr{R}_1}\right) \end{cases}$$

3) (a) $\dfrac{\partial \omega}{\partial I_2} + 1 > 0$; *and (b)* $\dfrac{\partial \omega}{\partial R_2} + 1 > 0$.

Proof. Properties 1) and 2) can be verified by directly differentiating (2.12) and noting that $\omega(I_2, R_2)$ is not differentiable at $I_2 + \epsilon R_2 = N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$ so that all inequalities are strict.

We now consider the three parts which compose property 3). We notice that they are trivially true for $I_2 + \epsilon R_2 > N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$, so we only need to consider $I_2 + \epsilon R_2 < N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$. By properties 1) and 2), we have

(a)
$$\frac{\partial \omega}{\partial I_2} + 1 = N\gamma_1 \left(\dfrac{1 + \dfrac{\beta_2 \omega(I_2, R_2)}{\beta_2 I_2 + N\sigma_1}}{\beta_2 I_2 + N(\gamma_1 + \sigma_1)} \right) > 0,$$

(b)
$$\frac{\partial \omega}{\partial R_2} + 1 = -\epsilon \left(\dfrac{\beta_2 I_2 + N\sigma_1}{\beta_2 I_2 + N(\gamma_1 + \sigma_1)} \right) + 1 = \dfrac{(1 - \epsilon)(\beta_2 I_2 + N\sigma_1) + N\gamma_1}{\beta_2 I_2 + N(\gamma_1 + \sigma_1)} > 0,$$

because $\omega(I_2, R_2) \geq 0$ by (2.12) and $1 - \epsilon \geq 0$ since $0 \leq \epsilon \leq 1$, and we are done.

We now prove Theorem 3.

Proof. Proof of 1: We consider the properties of the I_2- and R_2-nullcline in the region int(Λ). We have the following:

$$I_2' = 0 \implies f(I_2, R_2) = \beta_2 - \gamma_2 - \frac{\beta_2}{N}(\omega(I_2, R_2) + I_2 + R_2) = 0 \tag{C.1}$$

$$R_2' = 0 \implies g(I_2, R_2) = \gamma_2 I_2 - \sigma_2 R_2 - \frac{\beta_1}{N}(1 - \epsilon)\omega(I_2, R_2)R_2 = 0. \tag{C.2}$$

where $\omega(I_2, R_2)$ is given by (2.12). Also note that int(Λ) we have $I_2 > 0$ so that we have divided by I_2 in the I_2-nullcline. Due to the nonlinearities in I_2 and R_2 in (C.1) and (C.2), we will analyze $f(I_2, R_2)$ and $g(I_2, R_2)$ from implicit form.

In order to determine the R_2-intercepts, we note that $\mathscr{R}_{21} > 1$ implies $\beta_2\gamma_1(\beta_1 + \sigma_1) - \beta_1\gamma_2(\gamma_1 + \sigma_1) > 0$. We now evaluate $f(I_2, R_2)$ and $g(I_2, R_2)$ at $I_2 = 0$. This gives

$$f(0, R_2) = \beta_2 - \gamma_2 - \frac{\beta_2}{N}(\omega(0, R_2) + R_2) = 0$$
$$\implies \beta_2 - \gamma_2 - \frac{\beta_2}{N}\left(\frac{\sigma_1(N(\beta_1 - \gamma_1) - \epsilon\beta_1 R_2)}{\beta_1(\gamma_1 + \sigma_1)} + R_2\right) = 0$$
$$\implies R_2 = \frac{N(\beta_2\gamma_1(\beta_1 + \sigma_1) - \beta_1\gamma_2(\gamma_1 + \sigma_1))}{\beta_1\beta_2((1 - \epsilon)\sigma_1 + \gamma_1)} > 0$$
$$g(0, R_2) = -\left(\sigma_2 + \frac{\beta_1}{N}(1 - \epsilon)\omega(0, R_2)\right)R_2 = 0$$
$$\implies R_2 = 0$$

where we have utilized $1 - \epsilon \geq 0$ and $\omega(I_2, R_2) \geq 0$. It follows that, if $\mathscr{R}_{21} > 1$ is satisfied, then the I_2-nullcline has a positive R_2-intercept and the R_2-nullcline has the R_2-intercept $R_2 = 0$.

We now want to determine how the nullclines change as I_2 increases in the region $I_2 > 0$. We consider $R_2 = R_2(I_2)$ as a function of I_2 and implicitly differentiate to find $\dfrac{dR_2}{dI_2}$ on the surfaces $f(I_2, R_2)$ and $g(I_2, R_2)$. For the I_2-nullcline $f(I_2, R_2)$ (C.1), we have

$$\frac{d}{dI_2}f(I_2, R_2(I_2)) = -\frac{\beta_2}{N}\left(\frac{\partial\omega}{\partial I_2} + \frac{\partial\omega}{\partial R_2}\frac{dR_2}{dI_2} + 1 + \frac{dR_2}{dI_2}\right) = 0.$$

Solving for $\dfrac{dR_2}{dI_2}$ gives

$$\frac{dR_2}{dI_2} = -\left(\frac{\dfrac{\partial\omega}{\partial I_2} + 1}{\dfrac{\partial\omega}{\partial R_2} + 1}\right) < 0$$

by properties 3(a) and 3(b) of Lemma 1. It follows that the I_2-nullcline is strictly decreasing in int(Λ).

For the R_2-nullcline $g(I_2, R_2)$ (C.2) we have

$$\frac{d}{dI}g(I_2, R_2(I_2)) = \gamma_2 - \sigma_2\frac{dR_2}{dI_2} - \frac{\beta_1}{N}(1 - \epsilon)\left(\frac{\partial\omega}{\partial I_2}R_2 + \frac{\partial\omega}{\partial R_2}\frac{dR_2}{dI_2}R_2 + \omega(I_2, R_2)\frac{dR_2}{dI_2}\right) = 0.$$

Solving for $\dfrac{dR_2}{dI_2}$ gives

$$\frac{dR_2}{dI_2} = \frac{-\beta_1(1 - \epsilon)\dfrac{\partial\omega}{\partial I_2}R_2 + \gamma_2 N}{\beta_1(1 - \epsilon)\left(\dfrac{\partial\omega}{\partial R_2}R_2 + \omega(I_2, R_2)\right) + \sigma_2 N} \tag{C.3}$$

We now consider cases. If $\epsilon = 0$, then (C.3) reduces to

$$\frac{dR_2}{dI_2} = \frac{-\beta_1 \dfrac{\partial \omega}{\partial I_2} R_2 + \gamma_2 N}{\beta_1 \omega(I_2, R_2) + \sigma_2 N} \tag{C.4}$$

where we have note that $\dfrac{\partial \omega}{\partial R_2} = 0$ if $\epsilon = 0$ from property 2. of Lemma 1. Since $\omega(I_2, R_2) \geq 0$ we have $\beta_1 \omega(I_2, R_2) + \sigma_2 N > 0$, so we only need to consider the numerator of (C.4). We first note that (C.2) implies that

$$\gamma_2 = \frac{R_2}{I_2} \left(\sigma_2 + \frac{\beta_1}{N}(1 - \epsilon) R_2 \omega(I_2, R_2) \right). \tag{C.5}$$

Substituting (C.5) and the form of $\dfrac{\partial \omega}{\partial I_2}$ from property 1) of Lemma 1 into the numerator of (C.4) and factoring by $\omega(I_2, R_2)$ gives

$$\frac{R_2}{I_2} \left(\alpha_1 \omega(I_2, R_2) + \alpha_2 \right)$$

where

$$\alpha_1 = \frac{\beta_1(\sigma_1(\gamma_1 + \sigma_1)N^2 + 2I_2 N \beta_2 \sigma_1 + I_2^2 \beta_2^2)}{(\beta_2 I_2 + N\sigma_1)(\beta_2 I_2 + N(\gamma_1 + \sigma_1))} > 0$$

$$\alpha_2 = \frac{\sigma_2(\gamma_1 + \sigma_1)N^2 + (\beta_1\sigma_1 + \beta_2\sigma_2)I_2 N + I_2^2 \beta_1 \beta_2}{\beta_2 I_2 + N(\gamma_1 + \sigma_1)} > 0$$

It follows that, for the R_2-nullcline we have $\dfrac{dR_2}{dI_2} > 0$ if $\epsilon = 0$. For the case $\epsilon = 1$, we have that (C.3) reduces to

$$\frac{dR_2}{dI_2} = \frac{\gamma_2}{\sigma_2} > 0.$$

It follows that the R_2-nullcline is increasing in the region $\mathrm{int}(\Lambda)$ if $\epsilon = 0$ or $\epsilon = 1$.

We now consider the directions in the vector field for different regions of the planar state space (I_2, R_2). We show that $I_2' = f(I_2, R_2)$ and $R_2' = g(I_2, R_2)$ strictly decrease as R_2 increases. From (2.13), we have that

$$\frac{\partial}{\partial R_2} f(I_2, R_2) = \frac{\partial}{\partial R_2} \left[\frac{\beta_2}{N}(N - \omega(I_2, R_2) - I_2 - R_2)I_2 - \gamma_2 I_2 \right] = -\frac{\beta_2}{N} I_2 \left(\frac{d\omega}{dR_2} + 1 \right) < 0$$

by property 3(b) of Lemma 1. We also have

$$\frac{\partial}{\partial R_2} g(I_2, R_2) = \frac{\partial}{\partial R_2} \left[\gamma_2 I_2 - \sigma_2 R_2 - \frac{\beta_1}{N}(1 - \epsilon)\omega(I_2, R_2)R_2 \right]$$

$$= -\sigma_2 - \frac{\beta_1}{N}(1 - \epsilon) \left(\frac{d\omega}{dR_2} R_2 + \omega(I_2, R_2) \right).$$

$$= \begin{cases} -\sigma_2 - \dfrac{\beta_2}{N}\omega(I_2, R_2) < 0, & \text{if } \epsilon = 0 \\ -\sigma_2 < 0, & \text{if } \epsilon = 1 \end{cases}$$

where we have noted that $\dfrac{\partial \omega}{\partial R_2} = 0$ for $\epsilon = 0$ by property of 2) of Lemma 1. It follows that, if $\epsilon = 0$ or $\epsilon = 1$, we have that $I_2' < 0$ and $R_2' < 0$ if we are above both the I_2- and R_2-nullcline. The rest of the

cases follow similarly.

Proof of 2: Since the I_2-nullcline $f(I_2, R_2) = 0$ is strictly decreasing from a positive R_2-intercept, the R_2-nullcline $g(I_2, R_2) = 0$ is strictly increasing from an R_2-intercept of $R_2 = 0$, and solutions are restricted to Λ by Theorem 2, we have that the nullclines must have a unique intersection in int(Λ). Local exponential stability follows from Lemma 2 in Appendix D.

For global stability, we utilize the Poincaré-Bendixson Theorem [58]. Since the system is planar, it may not contain chaotic trajectories [58]. The only remaining possibilities are convergence to the unique steady state or convergence to a limit cycle. We will rule out the existence of a limit cycle using the modified Dulac's Criterion (Lemma 3 in Appendix C).

Consider the piecewise-defined vector field (2.13) and the function $\psi(I_2, R_2) = \dfrac{1}{I_2}$, which is continuous on int(Λ). We notice that (2.13) is continuous on int(Λ) and continuously differentiable everywhere on Λ except $I_2 + \epsilon R_2 = N\left(1 - \dfrac{1}{\mathscr{R}_1}\right)$.

We have that

$$
\frac{\partial}{\partial I_2}[\psi(I_2, R_2)f(I_2, R_2)] + \frac{\partial}{\partial R_2}[\psi(I_2, R_2)g(I_2, R_2)]
$$

$$
= \frac{\partial}{\partial I_2}\left[\frac{\beta_2}{N}(N - \omega(I_2, R_2) - I_2 - R_2) - \gamma_2\right] + \frac{\partial}{\partial R_2}\left[\gamma_2 - \left(\frac{N\sigma_2 + \beta_1(1 - \epsilon)\omega(I_2, R_2)}{NI_2}\right)R_2\right]
$$

$$
= -\frac{\beta_2}{N}\left(\frac{\partial \omega}{\partial I_2} + 1\right) - \frac{\sigma_2}{I_2} - \frac{\beta_1(1 - \epsilon)}{NI_2}\left(\frac{\partial \omega}{\partial R_2}R_2 + \omega(I_2, R_2)\right)
$$

$$
= \begin{cases} -\dfrac{\beta_2}{N}\left(\dfrac{\partial \omega}{\partial I_2} + 1\right) - \dfrac{\sigma_2}{I_2} - \dfrac{\beta_1}{NI_2}\omega(I_2, R_2) < 0, & \text{if } \epsilon = 0 \\[3mm] -\dfrac{\beta_2}{N}\left(\dfrac{\partial \omega}{\partial I_2} + 1\right) - \dfrac{\sigma_2}{I_2} < 0, & \text{if } \epsilon = 1 \end{cases}
$$

It follows by Lemma 3 in Appendix (D) that we cannot have a periodic orbit lying entirely within Λ. Consequently, from Poincaré-Bendixson Theorem $(\bar{I}_2, \bar{R}_2)$ is the global attractor for trajectories starting in int(Λ).

We now establish the location of the unique steady state $(\bar{I}_2, \bar{R}_2) \in$ int(Λ). We note that, since the I_2-nullcline is above the R_2-nullcline at $I_2 = 0$, for them to intersect in $0 < I_2 + \epsilon R_2 < N\left(1 - \dfrac{1}{\mathscr{R}_2}\right)$ it is sufficient for the R_2-nullcline to be above the I_2-nullcline along $\beta_2 I_2 + \beta_2 \epsilon R = N(\beta_2 - \gamma_2)$. Note that we have $\omega(I_2, R_2) = 0$ along this set. It consequently remains to consider the equations $f(I_2, R_2) = 0$ and $g(I_2, R_2) = 0$ along this set.

For the intersection of $f(I_2, R_2) = 0$ (C.1) and $\beta_2 I_2 + \beta_2 \epsilon R = N(\beta_2 - \gamma_2)$, we solve the following system of equations

$$
\begin{cases} \beta_1 I_2 & + & \beta_1 \epsilon R_2 & = & N(\beta_1 - \gamma_1) \\ \beta_2 I_2 & + & \beta_2 R_2 & = & N(\beta_2 - \gamma_2) \end{cases}
$$

to get

$$
R_f = \frac{N(\beta_2 \gamma_1 - \beta_1 \gamma_2)}{\beta_2 \beta_1 (1 - \epsilon)}.
$$

This is the value of R at the intersection of the I_2-nullcline and the transition line. For the intersection of $g(I_2, R_2) = 0$ (C.2) and $\beta_2 I_2 + \beta_2 \epsilon R = N(\beta_2 - \gamma_2)$, we solve the following system of equations

$$
\begin{cases}
\beta_1 I_2 & + & \beta_1 \epsilon R_2 & = & N(\beta_1 - \gamma_1) \\
\gamma_2 I_2 & - & \sigma_2 R_2 & = & 0
\end{cases}
$$

to get

$$
R_g = \frac{N\gamma_2(\beta_1 - \gamma_1)}{\beta_1(\sigma_2 + \epsilon\gamma_2)}.
$$

This is the value of R at the intersection of the R_2-nullcline and the transition line. In order to have $R_g > R_f$ we must have

$$
\beta_2\gamma_2(1 - \epsilon)(\beta_1 - \gamma_1) > (\sigma_2 + \epsilon\gamma_2)(\beta_2\gamma_1 - \beta_1\gamma_2) \implies \frac{\beta_1}{\gamma_1} > \frac{\beta_2}{\gamma_2}\left(\frac{\gamma_2 + \sigma_2}{(1 - \epsilon)\beta_2 + \epsilon\gamma_2 + \sigma_2}\right) \implies \frac{\mathscr{R}_1}{\mathscr{R}_2}\left(\frac{(1 - \epsilon)\beta_2 + \epsilon\gamma_2 + \sigma_2}{\gamma_2 + \sigma_2}\right) > 1.
$$

It follows that if $\mathscr{R}_{12} > 1$, then the steady state $(\bar{I}_2, \bar{R}_2)$ satisfies $0 < \bar{I}_2 + \epsilon\bar{R}_2 < N\left(1 - \frac{1}{\mathscr{R}_1}\right)$ which is sufficient to guarantee $\omega(\bar{I}_2, \bar{R}_2) > 0$ by (2.12). If $\mathscr{R}_{12} < 1$, however, the intersection will satisfy $\bar{I}_2 + \epsilon\bar{R}_2 > N\left(1 - \frac{1}{\mathscr{R}_1}\right)$ so that $\omega(\bar{I}_2, \bar{R}_2) = 0$ by (2.12), and we are done.

D. Supporting results

Lemma 2. *Consider the nonlinear planar system*

$$
\begin{cases}
\dfrac{dx}{dt} = f(x, y) \\
\dfrac{dy}{dt} = g(x, y).
\end{cases}
\tag{D.1}
$$

with a hyperbolic fixed point $(\bar{x}, \bar{y})$. Suppose that, at the fixed point $(\bar{x}, \bar{y})$, the x-nullcline is decreasing, the y-nullcline is increasing, and the region above $(\bar{x}, \bar{y})$ satisfies $x' < 0$ and $y' < 0$. Then $(\bar{x}, \bar{y})$ is locally exponentially stable.

Proof. Since $(\bar{x}, \bar{y})$ is a hyperbolic fixed point by assumption, the Hartman-Grobman Theorem [59, 60] guarantees the existence of a neighborhood around $(\bar{x}, \bar{y})$ where trajectories are mapped to a corresponding linear system with coefficient matrix given by the Jacobian of the nonlinear system evaluated at the fixed point. We furthermore have tangency of the nonlinear and nonlinear system nullclines at the fixed point and that $x' < 0$ and $y' < 0$ above the nullclines in the linear system. We now prove stability in the linearized system.

Consider the general linear system of ordinary differential equations in two variables:

$$
\begin{cases}
\dfrac{dx}{dt} = ax + by \\
\dfrac{dy}{dt} = cx + dy.
\end{cases}
\tag{D.2}
$$

In order to have $x' < 0$ above the x-nullcline, we require that:

$$x' < 0 \Longrightarrow ax + by < 0 \Longrightarrow by < -ax \Longrightarrow y > -\frac{a}{b}x.$$

This can only happen if $b < 0$. A similar argument for the y-nullcline gives that $d < 0$.

We now consider conditions 1 and 2. The x-nullcline is given by:

$$x' = 0 \Longrightarrow ax + by = 0 \Longrightarrow y = -\frac{a}{b}x.$$

In order for this to have a negative slope, we require that $-\frac{a}{b} < 0$. Given that $b < 0$, we must also have $a < 0$. Similarly, for the y-nullcline, we have:

$$y' = 0 \Longrightarrow cx + dy = 0 \Longrightarrow y = -\frac{c}{d}x.$$

For a positive slope, we require $-\frac{c}{d} > 0$ which, given that $d < 0$, implies that $c > 0$.

It follows that the coefficient matrix for the system has the indicated sign pattern

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} - & - \\ + & - \end{bmatrix}.$$

It immediately follows that:

$$\text{Tr}(A) = a + d = (-) + (-) < 0$$

and

$$\text{Det}(A) = ad - bc = (-)(-) - (-)(+) > 0.$$

It follows from classical dynamical systems theory that the steady state $(0,0)$ of the linear system (D.2) is exponentially stable. Local exponential stability of $(\bar{x}, \bar{y})$ for (D.1) follows, and we are done.

Lemma 3 (Modified Dulac's Criterion). *Consider the nonlinear planar system* (D.1) *and let Λ be a simply connected region in this plane. Consider a connected curve Γ lying within Λ which divides Λ into two regions, Λ_1 and Λ_2, (i.e., $\Lambda_1 \cup \Lambda_2 = \Lambda$ and $\Lambda_1 \cap \Lambda_2 = \Gamma$) both of which are simply connected. Suppose that f and g are continuous on Λ and continuously differentiable in the interiors of Λ_1 and Λ_2 (but potentially not continuously differentiable along Γ).*

Suppose there is a continuously differentiable function $\psi(x, y)$ such that

$$\frac{\partial}{\partial x}[\psi(x, y)f(x, y)] + \frac{\partial}{\partial y}[\psi(x, y)g(x, y)]$$

has a constant sign in the interior of Λ_1 and Λ_2. Then there are no nonconstant periodic orbits lying entirely in Λ.

Proof. Without loss of generality, we assume that

$$\frac{\partial}{\partial x}[\psi(x, y)f(x, y)] + \frac{\partial}{\partial y}[\psi(x, y)g(x, y)] > 0$$

in the interior of Λ_1 and Λ_2. Suppose there is a periodic orbit C enclosing a bounded region D. Without loss of generality, we assume that C is counter-clockwise oriented around D.

The Dulac criterion guarantees that C does not lie entirely in Λ_1 or Λ_2 so we need only consider the case where C crosses between Λ_1 and Λ_2. We note that C may cross Γ an arbitrary number of even times; however, each time it crosses and crosses back, we may generically divide the existing region into two subregions, both of which are simply connected. Consequently, it is sufficient to consider the first pair of crossings and induct to any subsequent pairs of crossings.

Suppose C crosses Γ and then crosses back. We define D_1 and D_2 to be the two open subregions of D created by this pair of crossings such that $\overline{D_1} \cup \overline{D_2} = D$ and $\overline{D_1} \cap \overline{D_2} \subseteq \Gamma$. We define C' to be a curve traversing the boundary between D_1 and D_2, i.e., $C' = \overline{D_1} \cap \overline{D_2}$. This curve can be oriented in either direction. We define $-C'$ to the curve oriented in the other direction. We now define C_1 and C_2 to be the two counter-clockwise oriented curves which traverse the boundary of D_1 and D_2 respectively. Note that these curves contain part of the curve C and one of either C' or $-C'$.

Now, since $f(x, y)$ and $g(x, y)$ are continuously differentiable within D_1 and D_2 we have that

$$\iint_{D_i} \frac{\partial}{\partial x}[\psi(x,y)f(x,y)] + \frac{\partial}{\partial y}[\psi(x,y)g(x,y)] \, dydx > 0$$

for $i = 1, 2$. By Green's Theorem, it follows that we have

$$
\begin{aligned}
0 &< \sum_{i=1}^{2} \iint_{D_i} \frac{\partial}{\partial x}[\psi(x,y)f(x,y)] + \frac{\partial}{\partial y}[\psi(x,y)g(x,y)] \, dydx \\
&= \sum_{i=1}^{2} \oint_{C_i} -\psi(x,y)g(x,y) \, dx + \psi(x,y)f(x,y) \, dy \\
&= \oint_{C} -\psi(x,y)g(x,y) \, dx + \psi(x,y)f(x,y) \, dy + \int_{C'} -\psi(x,y)g(x,y) \, dx + \psi(x,y)f(x,y) \, dy \\
&\quad + \int_{-C'} -\psi(x,y)g(x,y) \, dx + \psi(x,y)f(x,y) \, dy \\
&= \oint_{C} -\psi(x,y)g(x,y) \, dx + \psi(x,y)f(x,y) \, dy \\
&= \oint_{C} -\psi(x,y)\frac{dy}{dt} \, dx + \psi(x,y)\frac{dx}{dt} \, dy = 0
\end{aligned}
$$

This is a contradiction, so that we cannot have a periodic orbit which crosses Γ and then crosses back. We may now induct to multiple crossings by noting that each subsequent crossing will split one of the regions already considered in exactly the same way as the first split. In particular, it will be produce two simple connected subregions and the boundary between the split subregions may be parametrized by a connected curve. Since the periodic orbit C was chosen arbitrarily in Λ, we are done.