J. Math. Biol. (2009) 59:841–880
DOI 10.1007/s00285-009-0256-9

Mathematical Biology

An SIR epidemic model with partial temporary immunity modeled with delay

Michael L. Taylor · Thomas W. Carr

Received: 9 September 2008 / Revised: 13 February 2009 / Published online: 6 March 2009

Abstract The SIR epidemic model for disease dynamics considers recovered individuals to be permanently immune, while the SIS epidemic model considers recovered individuals to be immediately resusceptible. We study the case of temporary immunity in an SIR-based model with delayed coupling between the susceptible and removed classes, which results in a coupled set of delay differential equations. We find conditions for which the endemic steady state becomes unstable to periodic outbreaks. We then use analytical and numerical bifurcation analysis to describe how the severity and period of the outbreaks depend on the model parameters.

Keywords Epidemiology · Immunity · Resusceptible · Delay · Oscillations

Mathematics Subject Classification (2000) 34K13 · 34K18 · 34K25 · 34K26

1 Introduction

Compartmental models for viral and bacterial diseases separate a population into various classes based on the stages of infection (Anderson and May 1991). It is typical in the simpler compartmental models for the disease to either die out or approach an endemic, non-zero equilibrium. More complex temporal behavior can result from the inclusion of various pathological effects or other environmental factors; for example, the addition of seasonal forcing in the contact rate can cause the disease to exhibit recurrent epidemics or even become chaotic. In this paper, we consider the role that temporary immunity plays in the spread of diseases such as cholera, pertussis,

M. L. Taylor · T. W. Carr (✉)
Department of Mathematics, Southern Methodist University,
Dallas, TX 75275-0156, USA
e-mail: tcarr@smu.edu

influenza and malaria. We construct a system of delay differential equations (DDEs) as a model for diseases that exhibit temporary immunity for a fraction of the recovered individuals. When the fraction of individuals who become resusceptible is small, the system will evolve to either the disease-free steady state or the endemic state depending upon the other system parameters. For higher values of the resusceptible fraction, there is a Hopf bifurcation to oscillatory solutions that indicates recurrent epidemics of the diseased population. In this paper, we identify the specific conditions required for oscillatory solutions and use asymptotic methods to explicitly construct and then investigate the recurring epidemics.

One of the most fundamental compartment models based on differential equations is the SIR model described by Eqs. 1 below (Anderson and May 1991). This model classifies individuals to be susceptible (S), infectious (I), or removed (R) and permanently immune, and is appropriate for diseases such as measles and mumps. Individuals are born into the susceptible class, and after having the disease, become part of the removed class. At the population level, there are disease-free and endemic steady states, where the parameters determine which is stable:

$$\begin{aligned} \frac{dS}{dt} &= b - \beta SI - \mu S, \\ \frac{dI}{dt} &= \beta SI - (\mu + \gamma)I, \\ \frac{dR}{dt} &= \gamma I - \mu R. \end{aligned} \tag{1}$$

The total population size has been normalized to one and S, I and R represent the fraction of the total population in each compartment. b and μ are birth and death rates, respectively, and β is the transmission coefficient related to the number of contacts that successfully transmit disease. γ is the recovery rate such that $1/\gamma$ is the mean time of infection; $dR/dt \sim +\gamma I$ is the rate at which individuals recover and become immune.

Other compartment combinations may more accurately model other diseases. For example, an SI model describes a disease with two stages such as herpes or HIV, where individuals are infectious for life and never removed. An SIS model describes the case when individuals recover from the disease but there is no immunity, and they return to the susceptible class. Examples include sexually transmitted diseases, plague and meningitis. Finally, an SEIR model includes an "exposed" class of individuals who are not yet infectious, and is appropriate for yellow fever. Additional "forces" that may be included in the model are disease-related death, vaccination, and seasonal variations in infectiousness, to name just a few.

In this paper we will use an SIRS model to investigate the effect of temporary immunity on the prevalence of a disease in a population; that is, when the removed individuals eventually return to the susceptible class (see Eqs. 2). As mentioned above, temporary immunity plays a role in the spread of many human diseases. In cholera and pertussis, for example, immunity weakens over time such that a fraction of recovered individuals becomes resusceptible to infection, typically after a year or longer. Influenza and malaria are characterized by multiple strains and high mutability within each

strain. As a result, individuals recover with effectively temporary immunity, resistant to reinfection from a specific strain but susceptible to infection from other strains. We will consider temporary immunity to have a fixed duration of time, and modeled by a delay term in the susceptible and recovered population equations; the SIRS equations then become a system of DDEs.

We will find that the duration of temporary immunity or delay time plays a critical role in determining both the onset and properties of recurrent epidemics. For example, there is a minimum delay time such that if the duration of temporary immunity is less than the required minimum time, then sustained epidemics will not occur. For some intervals of the delay time, the resusceptible fraction required to generate recurrent oscillations is very small. The delay time determines whether the appearance of recurrent epidemics will be via a supercritical Hopf bifurcation and hence the population will exhibit small epidemics, or if the Hopf bifurcation will be subcritical and the population will experience large pulsating epidemics. In general, our analysis and numerical simulations will describe how the amplitude and period of recurrent epidemics depend upon the resusceptible fraction and the delay time, as well as the other system parameters.

In the remainder of the introduction, we provide an overview of how DDEs have been used to model other disease characteristics; we also discuss other alternatives for modeling temporary immunity. Later sections are generally either focused on epidemiological considerations and results, or on mathematical analysis. Thus, readers whose interests are geared more towards modeling and the general effects of modeling temporary immunity using DDEs will want to focus on the next two sections and then proceed to the final discussion section. More specifically, in Sect. 2 we introduce the specific model that we will analyze and derive a non-dimensionalized form that will serve as the basis for our analysis. In Sect. 3.1 we derive the conditions for the appearance of recurrent epidemics (i.e., oscillations). We follow that with the results of numerical simulations that provide an overview of the type of behavior the model can exhibit under different parameter conditions. In the final discussion section we summarize the mathematical results and analysis and conclude by returning to discuss the biological interpretation and ramifications of our results.

Readers who are interested in the application of singular perturbation methods to DDEs will want to look at the analysis in Sects. 4–6. In the first two, we use the method of multiple scales, modified to account for delay, to analyze small amplitude epidemics local to a Hopf bifurcation point. In the third, we derive an iterated map to describe large-amplitude pulsating epidemics that occur for high resusceptibility.

1.1 Epidemic models with temporary immunity and delays

Differential equations with delays have been used to examine the effect of disease characteristics such as a fixed latency or infectious period (Cooke and Van Den Driessche 1996) and a period of temporary immunity (Hethcote et al. 1981). Delays have also been used to account for characteristics of the host population such as maturation time (Cooke et al. 1999) or newborn immunity (Hethcote 2000). In this section, we briefly review some of these uses of delays in various disease models.

Upon infection, it is often the case that an individual is not immediately infectious. During this latency period, the host remains part of the exposed class; that is, no longer in the susceptible class but not yet in the infectious class. For diseases with a constant latency period, incorporating a delay represents the constant period of latency more accurately than a separate exposed class. A fixed recovery or infectious time may be modeled similarly. For an SEIRS model with fixed latent and recovery times, Cooke and Van Den Driessche (1996) introduced two delays. The first delay, representing a fixed latency period, appears in the transmission term of the infectious population equation:

$$\frac{dI}{dt} = \beta S(t)I(t) \quad \text{becomes} \quad \frac{dI}{dt} = \beta S(t-\hat{\tau})I(t-\hat{\tau})e^{-\mu\hat{\tau}}.$$

The change in the number of infectious individuals at time t is due to those who were exposed at a prior time $(t - \hat{\tau})$. Similarly, the second delay is used to account for a fixed recovery time.

In humans, a newborn inherits its mother's antibodies, which help to provide additional protection while the infant's immune system continues to develop. A separate class can be added for newborns with passive immunity from maternal antibodies (Hethcote 2000). Alternatively, the model may incorporate a delay representing the time separating birth from the initial vaccinations.

Another characteristic of many species is a maturation period prior to reproductive adulthood. In humans, children become mature and capable of reproduction in their early to mid-teens. A constant time prior to maturation can lead to a delayed model as discussed in Cooke et al. (1999). In fact, one could separate the human maturation process into several stages of development including infancy, childhood, and adolescence. As long as each stage has equal survival rates, Cooke's model allows for several stages of growth preceding reproductive adulthood to be treated as one long maturation delay.

Temporary immunity occurs in a number of diseases, as mentioned in the introduction. The fixed recovery time and corresponding delay in Cooke and Van Den Driessche (1996) represent temporary immunity. Hethcote et al. (1981) incorporates a delay term into an integro-differential SIRS model to represent temporary immunity. The resulting system reduces to a single first-order integro-differential equation as the S class decouples $(S(t) = 1 - I(t))$ and the $R(t)$ population gets absorbed by the delay term. As an alternate approach, Hethcote also suggests an $SIR_1R_2, \ldots, R_n - S$ model to delay the return of individuals to the susceptible class. In other words, rather than solving a delayed system, he looks at a system with multiple recovered classes and finds that $n \geq 3$ has qualitatively the same dynamics as the delayed system.

We will consider the time for temporary immunity to be a fixed non-zero constant that is the same throughout the population, which we will model using DDEs. As described above, an alternative is to use multicompartment ODE-based models. An even more general approach is to allow for a distribution of delay times such that the model becomes an integro-differential equation, where a kernel function may be designed to model a specific distribution of immune times (Brauer and Castillo-Chavez 2001). For example, Diekmann and Montijn (1982) (see also Chow et al. 1985)

considered a fixed period of temporary immunity as part of an age-structured integral equation with a time-dependent force of infection. They obtain a transcendental characteristic equation for the stability of the endemic disease state that is similar to the one we study in Sect. 3.1. Similar to our model, the endemic state becomes unstable to periodic oscillations corresponding to recurrent epidemics in the population (Chow et al. 1985). Broadly speaking, using fixed delays can be considered more general than the ordinary differential equation approach in that we do not allow anyone to be immediately resusceptible. However, while less general than an integro-differential model, using DDEs with fixed delays will often be easier to analyze.

2 SIR epidemic model with partial-temporary immunity

2.1 SIR model

A simple model that accounts for temporary immunity is the SIRS model given by:

$$\begin{aligned} \frac{dS}{dt} &= b - \beta SI - \mu S + \sigma R, \\ \frac{dI}{dt} &= \beta SI - (\mu + \gamma)I, \\ \frac{dR}{dt} &= \gamma I - (\mu + \sigma)R. \end{aligned} \tag{2}$$

The coefficient $1/\sigma$ is the mean time of immunity and $dS/dt \sim +\sigma R$ is the rate at which individuals again become susceptible. More specifically, immunity times for individuals range from zero (i.e., no immunity and immediately resusceptible) to infinite (i.e., permanent immunity) according to an exponential distribution of immunity times, where the mean immunity time is $1/\sigma$ (Brauer and Castillo-Chavez 2001).

Our goal is to investigate the effect of a fixed duration of immunity; that is, when an individual recovers they are immune for a fixed duration $\hat{\tau}$, at which time they become resusceptible. Thus, we start with an SIR model and couple the R class and S class as in the SIRS model. However, the coupling term will contain a fixed delay such that $dR/dt \sim -\gamma I(t - \hat{\tau})$ and $dS/dt \sim +\gamma I(t - \hat{\tau})$. This indicates that individuals who become resusceptible at time t had entered the R class at time $t - \hat{\tau}$. Thus, they have been immune for a duration $\hat{\tau}$.

We must also take into account the survival rate of recovered individuals. For example, consider a population $N(t)$ with only a death process,

$$\frac{dN(t)}{dt} = -\mu N(t).$$

Solving for $N(t)$ gives,

$$N(t) = N(0)e^{-\mu t}, \quad \text{for } t \geq 0.$$

Thus, the fraction of the original population who survive from time $t = 0$ to time t is $e^{-\mu t}$. In our model, $e^{-\mu\hat{\tau}}$ is the fraction of individuals who recover at time $t - \hat{\tau}$ who survive to time t (Brauer and Castillo-Chavez 2001).

Finally, we allow that only a fraction r_γ of the population might become resusceptible, while the remaining fraction, $1 - r_\gamma$, remain permanently immune; we shall refer to r_γ as the "resusceptible fraction". The model we consider is then given by:

$$\frac{dS}{dt} = \mu[1 - S(t)] - \beta I(t)S(t) + r_\gamma \gamma e^{-\mu\hat{\tau}} I(t - \hat{\tau}), \tag{3}$$

$$\frac{dI}{dt} = \beta I(t)S(t) - (\mu + \gamma)I(t), \tag{4}$$

$$\frac{dR}{dt} = \gamma I(t) - \mu R(t) - r_\gamma \gamma e^{-\mu\hat{\tau}} I(t - \hat{\tau}). \tag{5}$$

To simplify calculations, our model considers equal birth and death rates ($b = \mu$). Thus, the population size is fixed and normalized to $N = 1$ so that by summing the three equations we have $R(t) = 1 - S(t) - I(t)$. In further analysis it then suffices to consider only Eqs. 3 and 4, where R is determined by the above constraint. Finally, we note that our model is similar to that considered in Brauer and Castillo-Chavez (2001) (see Sect. 7.7) except that we include the partial population immunity with $0 \leq r_\gamma \leq 1$ and our analysis will retain the effects of the birth and death terms.

2.2 Steady states and non-dimensionalization

Equations 3 and 4 have two steady states, a disease-free steady state ($S_c = 1$, $I_c = 0$) valid for all parameter values and an endemic steady state:

$$S_c = \frac{1}{\mathcal{R}_0}, \quad I_c = \frac{\frac{\mu}{\beta}(\mathcal{R}_0 - 1)}{1 - \frac{r_\gamma \gamma}{\mu+\gamma} e^{-\mu\hat{\tau}}}, \qquad \text{where } \mathcal{R}_0 = \frac{\beta}{\mu + \gamma}. \tag{6}$$

$\mathcal{R}_0$ is the basic reproductive number (Anderson and May 1991) and determines whether the disease dies out or persists in a population. Specifically, from a linear stability analysis of Eqs. 3 and 4, we find that for $\mathcal{R}_0 < 1$ the disease-free steady state is stable, while for $\mathcal{R}_0 > 1$ the disease-free steady state is unstable. The endemic steady state exists only if $I_c > 0$ and hence $\mathcal{R}_0 > 1$; its stability will depend upon r_γ as we will describe in the next section. In addition, using energy arguments similar to those in Pieroux and Erneux (1996); Carr et al. (2000) applied to the rescaled system Eqs. 9 (derived below), it is possible to show that for small values of r_γ the endemic state is globally stable for $\mathcal{R}_0 > 1$.

To simplify further analysis, we define new variables for the deviations from the non-zero endemic state and rescale time:

$$I = I_c(1 + y), \quad S = S_c\left(1 + \sqrt{\frac{I_c}{S_c}}x\right), \tag{7}$$

$$\text{and} \quad s = \beta\sqrt{S_c I_c}\, t, \quad \text{then let } s \to t \tag{8}$$

Substituting Eqs. 7 and 8 into Eqs. 3 and 4 results in:

$$\begin{aligned} \frac{dx}{dt} &= -y - \epsilon x(a + by) + ry(t - \tau), \\ \frac{dy}{dt} &= x(1 + y), \end{aligned} \tag{9}$$

where:

$$\epsilon = \sqrt{\frac{\mu\beta}{\gamma^2}} \ll 1. \tag{10}$$

$$\epsilon a = \frac{\mu + \beta I_c}{\beta\sqrt{S_c I_c}}, \quad \epsilon b = \sqrt{\frac{I_c}{S_c}}, \quad r = \frac{r_\gamma \gamma}{\mu + \gamma} e^{-\mu\hat{\tau}}, \quad \tau = \beta\sqrt{S_c I_c}\,\hat{\tau} \tag{11}$$

and where a and b are taken to be $O(1)$. Recall that $1/\mu$ is the mean lifetime of individuals who die a natural death; thus, to have $\epsilon \ll 1$ we require that individuals be long lived, $\mu \ll 1$, relative to the disease rate constants β and γ. Additional understanding can be gained by considering ϵ rewritten as

$$\epsilon = \sqrt{\left(\frac{\frac{1}{\gamma}}{\frac{1}{\mu}}\right)\left(\beta\frac{1}{\gamma}\right)} \ll 1. \tag{12}$$

The second term, $(\beta\frac{1}{\gamma})$, is effectively $\mathcal{R}_0$ (see Eq. 6 with $\mu \ll 1$) and is the number of infections per infected individual. The first term is the mean infectious time divided by the mean lifetime. Thus, ϵ is related to the number of infections per infected multiplied by the fraction of an individual's lifetime during which they will be infectious. Smaller ϵ implies a shorter and/or weaker infection.

The rescaled resusceptible fraction r will be the primary control parameter that we use to study the effect of temporary immunity. Its physical interpretation is as follows: $\exp(-\mu\hat{\tau})$ is the fraction of those who recover who survive to time $\hat{\tau}$. Thus, $r_\gamma \exp(-\mu\hat{\tau})$ is the faction of those who recover who ultimately become resusceptible. $\gamma/(\mu + \gamma)$ can be considered to be the fraction of those individuals who become infectious and then recover, the others leaving the population via natural death. Thus, taken altogether, r represents the fraction of those in the susceptible class who return to the susceptible class after being infectious.

If all individuals who recover become resusceptible, then the original, unscaled, resusceptible fraction is $r_\gamma = 1$. From Eq. 11 we then have that $0 \le r \le r_{\max} < 1$, where:

$$r_{\max} = \frac{\gamma}{\mu + \gamma} e^{-\mu\hat{\tau}}. \tag{13}$$

The reason that $r_{\max} < 1$ is, as described in the previous paragraph, because some individuals leave the population due to natural death ($\mu \neq 0$). For the scaling assumptions given above with $\gamma \gg \mu$ and $\mu \ll 1$, then for $\hat{\tau} = O(1)$, $r_{\max}$ is slightly less

than one. However, for delays that are on the order of the lifetime of the individual such that $\hat{\tau} = O(1/\mu)$, then $r_{\max} \approx 1/e$.

Finally, we note that for $r = 0$, Eqs. (9) are a weakly damped conservative system that has been studied both as a model for disease transmission and also lasers (Schwartz and Smith 1983; Schwartz and Erneux 1994; Kim et al. 2005). Indeed, in the context of laser systems under the influence of delay, Eqs. 9 have been studied extensively in Pieroux et al. (1994, 2000); Pieroux and Erneux (1996); Bestehorn et al. (2000). We will be able to adopt results from these previous works on lasers to understand some of the dynamics that we observe for disease transmission. However, there is an important difference between our work and the previous work on lasers. Specifically, our feedback parameter r is positive and can be $O(1)$, whereas in the laser studies, the feedback is negative and small. For weak feedback when, as we shall see, the delay induces small-harmonic oscillations, the sign on r is not consequential, and this is the regime where we can borrow results from the earlier works. However, if $r = O(1)$, then the delay term is larger than the $O(\epsilon)$ dissipation such that $dx/dt \approx -y + ry(t - \tau)$. This is when we will need new analysis to describe the harmonic and pulsating oscillations that are observed.

3 Periodic outbreaks due to delay

3.1 Linear stability of the endemic state

In this section we analyze the stability of the non-zero endemic state, which in the new variables is given by $(x, y) = (0, 0)$. For $\mathcal{R}_0 > 1$, the dynamics of the system near the endemic steady state can be approximated by linearizing the system near $x = y = 0$. As with constant-coefficient ordinary differential equations, we look for solutions proportional to $e^{\lambda t}$, which results in the following characteristic equation for λ:

$$\lambda^2 + \epsilon a\lambda + 1 - re^{-\lambda\tau} = 0. \tag{14}$$

For the equilibrium to be asymptotically stable, all of the eigenvalues must have negative real parts. Unlike a system of ODEs with a polynomial characteristic equation, the characteristic equation for this system of DDEs is a transcendental equation in λ, whose solution presents both analytical and numerical challenges. Before solving for λ in general, we consider two limiting cases that lend insight into the linear stability of the non-zero steady state.

Permanent immunity When $r = 0$ the delay term is removed from the model so that hosts recover with permanent immunity. Equations 9 are then a scaled version of the SIR model, Eqs. 1, and the endemic steady state, given by $(x, y) = (0, 0)$, is stable for $\mathcal{R}_0 > 1$.

No delay When $\tau = 0$ there is no delay between recovery and reentering the susceptible class; this is equivalent to an SIS model. However, for $r < 1$ only a fraction of

the population becomes resusceptible, while the remaining are permanently immune. We find that for $\tau = 0$ and $r < 1$ the endemic state is stable.

We now return to analyzing Eq. 14 in general. When hosts recover with partial temporary immunity ($r \neq 0$, $\tau \neq 0$) we find that the non-zero endemic state becomes unstable to periodic solutions through a Hopf bifurcation. To determine when the Hopf bifurcation occurs, we let $\lambda = i\omega$ in Eq. 14 and obtain

$$0 = (1 - \omega^2) \tan \omega\tau + \epsilon a \omega, \tag{15}$$

$$r_h^2 = \epsilon^2 a^2 \omega^2 + (1 - \omega^2)^2, \tag{16}$$

where $r = r_h$ is the value of r at the Hopf bifurcation point. Equation 15 is a transcendental equation for the frequency, ω, of the periodic solutions that emerge from the Hopf bifurcation. Given ω, the value of r_h at the Hopf bifurcation point can be determined by Eq. 16.

In general, Eq. 15 must be solved numerically. However, because $\epsilon \ll 1$ then we expect solutions $\omega \approx 1$ or $\omega\tau \approx m\pi$, m an integer. For the case $\omega \approx 1$ we let $\omega = 1 + \epsilon\omega_1 + O(\epsilon^2)$, substitute into Eq. 15 and find that

$$\omega = 1 + \epsilon \frac{1}{2} a \cot \tau + O(\epsilon^2). \tag{17}$$

Using this result in Eq. 16 we find

$$r_h = -\epsilon \frac{a}{\sin \tau} + O(\epsilon^2). \tag{18}$$

So that r_h is positive we require that $\sin \tau < 0$, which implies that this approximation is valid for values of the delay in the intervals $\tau \in (\pi, 2\pi), (3\pi, 4\pi), \ldots$. We shall refer to the oscillations that appear via the Hopf bifurcation with $\omega \approx 1$ as the *natural mode* because $\omega \approx 1$ is the natural quasifrequency of the system without delay; that is, without delay, perturbations from the endemic state oscillate and decay with quasifrequency $\omega \approx 1$.

For the case $\omega \approx m\pi/\tau$ we let $\omega = m\pi/\tau + \epsilon\omega_1 + O(\epsilon^2)$ and find that

$$\omega = \frac{m\pi}{\tau} - \epsilon \frac{am\pi}{\tau^2 - (m\pi)^2} + O(\epsilon^2), \tag{19}$$

and for r we have that

$$r_h = \pm \left(\left[1 - \left(\frac{m\pi}{\tau} \right)^2 \right] + \epsilon \frac{2a(m\pi)^2}{\tau(\tau^2 - (m\pi)^2)} \right) + O(\epsilon^2), \tag{20}$$

where the positive solution is used if m is even and the negative solution is used if m is odd [the sign is determined by examining the original real and imaginary parts of Eq. 14 expressed in terms of $\cos(\omega\tau)$ and $\sin(\omega\tau)$]. We shall refer to the oscillations that appear with ω given by Eq. 19 as *delay modes* because their frequency is determined by the delay time.

We note that Eqs. 15 and 16 are essentially equivalent to Eqs. 3.1 and 3.2 of Pieroux et al. (1994). Thus, Eqs. 17 and 18 for the natural mode also appear and are studied in Pieroux et al. (1994). However, because Pieroux et al. restrict themselves to $r \ll 1$, they do not consider the delay modes described by Eqs. 19 and 20.

In Fig. 1 we plot $r_h(\tau)$ and $\omega(\tau)$ as a function of the delay τ. The (+) are the result of numerically evaluating Eqs. 15 and 16, while the solid curves are the analytical approximations derived above. In Fig. 1b we see that in the intervals $\tau \in (\pi, 2\pi), (3\pi, 4\pi), \ldots$ the largest value of r such that the endemic state remains stable is given by Eq. 18 such that $r_h = O(\epsilon)$, the frequency of the resulting oscillations is $\omega \sim 1$.

In the intervals $\tau \in (0, \pi), (2\pi, 3\pi), \ldots$ the first instability is one of the delay modes with r_h given by Eq. 20; these are indicated by the light solid curves in Fig. 1. In these intervals, the delay time determines the critical value of the resusceptible fraction for a Hopf bifurcation. Similarly, the frequency is locked to some multiple of the inverse delay time, as described by Eq. 19.

For any value of the delay, as r is increased beyond the least value of r_h, additional delay modes bifurcate when r_h satisfies Eq. 20. These bifurcations will not be observed directly because the steady state has already become unstable. However, because they indicate that more oscillatory modes are unstable, they contribute to more complex system behavior when $r = O(1)$. In addition, they are the origin of the multi-stability of oscillatory solutions that is described in Pieroux et al. (1994).

In the left-most interval of delay times when $\tau \in (0, \pi)$, the value of m is 1 so that $r \approx -1 + (\pi/\tau)^2$. Because we have the restriction $r < r_{\max}$, there is a minimum

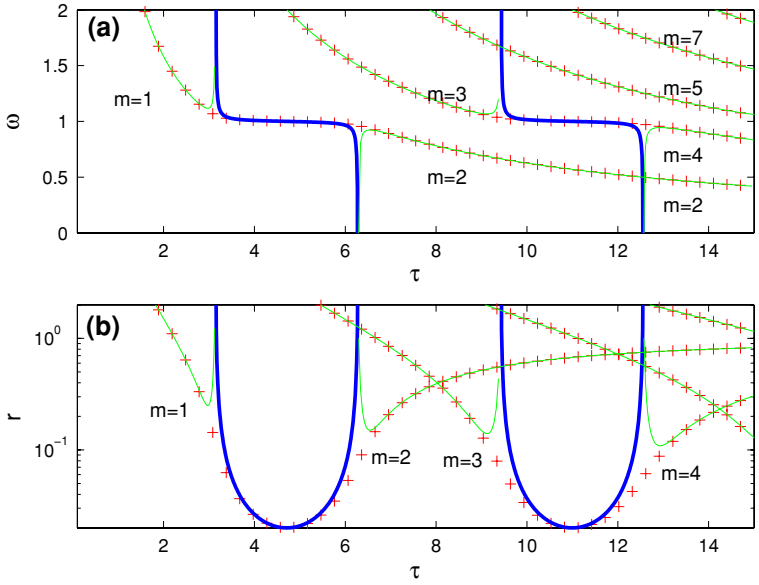

Fig. 1 Analytical and numerical solutions (for ω and r) of the characteristic equation (Eq. 14) with $\epsilon = 0.014$ and $a = 1.41$. In **a** and **b**, *solid dark* and *light lines* are analytical approximations for $\omega \approx 1$ and $\omega \approx m\pi/\tau$ respectively. *Crosses* are for numerical approximations

value of the delay such that a Hopf bifurcation will be observed. Specifically, there will only be a physically realizable Hopf bifurcation if $\tau > \tau_{\min}$, where $\tau_{\min}$ satisfies

$$r_{\max}(\tau_{\min}) = -1 + \left(\frac{\pi}{\tau_{\min}}\right)^2, \tag{21}$$

and $r_{\max}$ is given by Eq. 13. A weaker but simpler estimate can be obtained by setting $r_{\max} = 1$

$$\tau_{\min} \approx \sqrt{\frac{\pi}{2}}. \tag{22}$$

For the parameter values used to generate Fig. 1, we have found that Eq. 22 is a good estimate for the minimum value of the delay needed to result in a Hopf bifurcation of the endemic state. In terms of the original variables we have that to first approximation

$$\hat{\tau}_{\min} \approx \sqrt{\frac{\pi}{2}\frac{1}{\mu\gamma}\left(\frac{1-r_\gamma}{\mathcal{R}_0 - 1}\right)}. \tag{23}$$

We see that $\hat{\tau}_{\min}$ decreases as $r_\gamma \to 1$ or as $\mathcal{R}_0$ increases. Thus, the higher the resusceptible fraction or the more virulent the disease, the more sensitive the population is to oscillations due to delays.

As can be seen in Fig. 1, the analytical approximations are singular when $\tau = m\pi$. A more refined perturbation series expansion should resolve the singularities. However, we will not pursue that at this time and instead proceed to examine the periodic outbreaks that occur following the Hopf bifurcation of the natural mode when $\omega \approx 1$.

3.2 Numerical simulations of periodic outbreaks

In this section we present numerical and analytical results for the effect of temporary immunity. We will organize our discussion of the results by considering the delay time τ fixed, and observing the system's output as the resusceptible fraction r is increased from 0 to $r_{\max}$. We will present analytical approximations to the numerical results but postpone their derivation until later sections. To begin, we first summarize the system's behavior. We assume that $\mathcal{R}_0 \sim \beta/\gamma > 1$ such that the endemic steady state is stable; thus, Eqs. 9 exhibit damped oscillations to the endemic steady state $(x, y) = (0, 0)$. As the resusceptible fraction r is increased there is a Hopf bifurcation such that the endemic state becomes unstable. After the Hopf bifurcation, as r is increased towards $r_{\max}$, the system exhibits periodic oscillations corresponding to recurrent epidemics, which increase from being small amplitude and nearly harmonic to large amplitude and pulsating.

3.2.1 Parameter values

For most of our simulations the chosen parameter values ($a = 1.41$, $b = 0.71$, and $\epsilon = 0.014$) correspond to $\gamma \approx 100$, $\beta \approx 200$, and $\mu = 0.01$. If time is measured in years, these parameters correspond to a mean lifetime of 100 years, a mean recovery time on the order of a week, and an $\mathcal{R}_0 \sim \beta/\gamma$ of approximately 2. These values are roughly appropriate for a wide variety of human diseases (see Anderson and May 1991, Tables 3.1 and 4.1). More generally, our results maintain fidelity under the criteria $\epsilon \ll 1$. Thus, we will also show results for higher values of the transmission coefficient β corresponding to $\mathcal{R}_0 \approx 8$ and $\mathcal{R}_0 \approx 14$.

For most of our simulations we use a delay $\tau = 3\pi/2$, which in real-time units corresponds to a temporary immunity time in the range of 5–10 years. (the interval in the real-time delay is due to the scaling in Eq. 8 that depends on the bifurcation parameter). We will also consider immunity times that are both longer and shorter.

3.2.2 Time evolution

Using the numerical routine DDE_SOLVER (Thompson and Shampine 2006), we have computed solutions to Eq. 9 for various values of r. The results are shown in Fig. 2 for $r = 0.005$, 0.02, 0.03, and 0.9. In Fig. 2a the fraction of recovered individuals that become resusceptible is very small ($r \ll 1$) and the system exhibits oscillations that decay to the endemic state $(x, y) = (0, 0)$. When $r \approx 0.03$ the endemic state becomes unstable via a Hopf bifurcation. Increasing r corresponds to increasing the fraction of recovered individuals that are re-injected into the susceptible population. For $r > r_h$ the greater influx of susceptible individuals sustains recurring outbreaks as seen by the harmonic oscillations in Fig. 2b. Further increases in the fraction of individuals who become resusceptible drive the system to generate pulsating epidemics, Fig. 2c and d.

3.2.3 Bifurcation diagrams

Figure 3 are numerical bifurcation diagrams generated using Matlab routine DDE_BIFTOOL (Engelborghs et al. 2001; Luzyanina et al. 2005). They illustrate the relation between the resusceptible fraction and the period and amplitude of solutions. Specifically, we see two important regions of change. The first is immediately following the Hopf bifurcation, where the amplitude and frequency of the oscillatory solutions increase sharply. In this regime the oscillatory solutions are nearly harmonic as in Fig. 2b. The second region is for values of r greater than approximately 0.2, where the amplitude and period increase more gradually. In this latter regime the oscillations become pulsating as shown in Fig. 2c and d. Each of these regions will be described by separate asymptotic approximations in later sections. Finally, the oscillatory solutions are stable throughout the full range of the parameter r.

In Fig. 4 we focus on the vicinity of the Hopf bifurcation point for three different values of the delay. In each case, the delay is in the interval $\tau \in (\pi, 2\pi)$ such that the Hopf bifurcation is the natural mode described by Eqs. 17 and 18. When the delay is

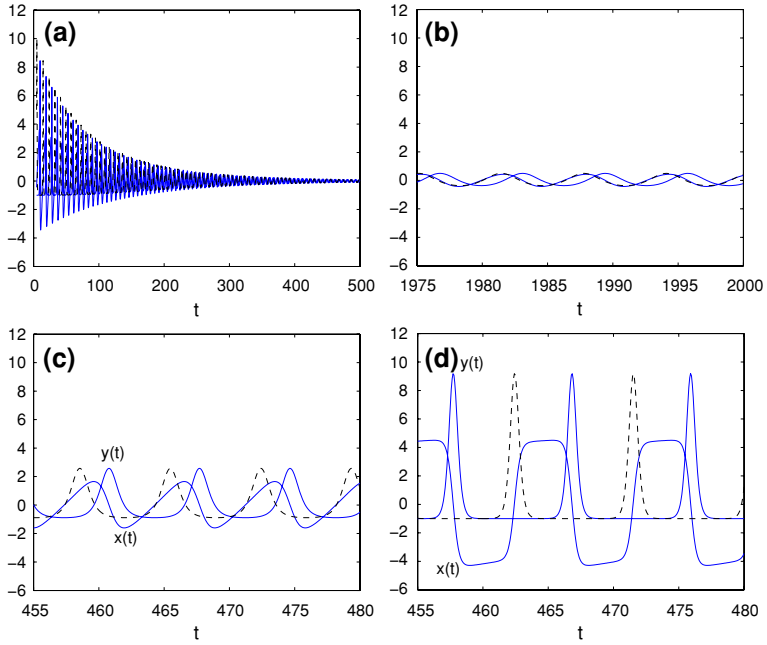

Fig. 2 Periodic solutions from DDE_SOLVER with $a = 1.41$, $b = 0.71$, and $\epsilon = 0.014$, and $\tau = 3\pi/2$. **a** $r = 0.005$, **b** $r = 0.02$, **c** $r = 0.03$, and **d** $r = 0.9$. Periodic solutions appear as r increases beyond the Hopf bifurcation point $r = r_h$. As r increases towards 1, $y(t)$ becomes pulsating, while $x(t)$ transitions from **b** harmonic to **c** *triangular waves* then **d** *square waves*. The *dashed curve* corresponds to $y(t - \tau)$

$\tau = 3\pi/2$, this corresponds to the minimum of the neutral stability curve in Fig. 1. For values of the delay to the left of the minimum, the bifurcation is supercritical. However, for values of the delay greater than the minimum when $\tau > 3\pi/2$, the bifurcation is subcritical; in this case, there is a region of bistability between the endemic state and oscillatory (and pulsating) solutions.

In Fig. 5a we examine values of the delay to the left of the minimum at $\tau = 3\pi/2$. When $\tau = 4.2$ the Hopf bifurcation point is given by Eq. 18; the dashed curve is the analytical prediction, derived in Sect. 4, given by

$$r = -\frac{\epsilon a}{\sin\tau}\left[1 + B^2\left(\frac{1}{6}\tau\cot\tau + \frac{5}{18} - \frac{4}{9}\cos\tau\right)\right], \tag{24}$$

where B is proportional to the amplitude of the oscillations. More generally, by solving for the amplitude we have an explicit expression $B = B(r, \epsilon a, \tau)$ for how the amplitude of the epidemics depends upon the diseases parameters. Notice that the direction of the bifurcation is determined by the term $1/\sin\tau$ and, hence, controlled by the time duration that individuals are temporarily immune.

When $\tau = 2.8$ the bifurcation is that of the delay mode with $m = 1$, with the bifurcation point given by Eq. 20. The dashed curve is our analytical prediction, derived in

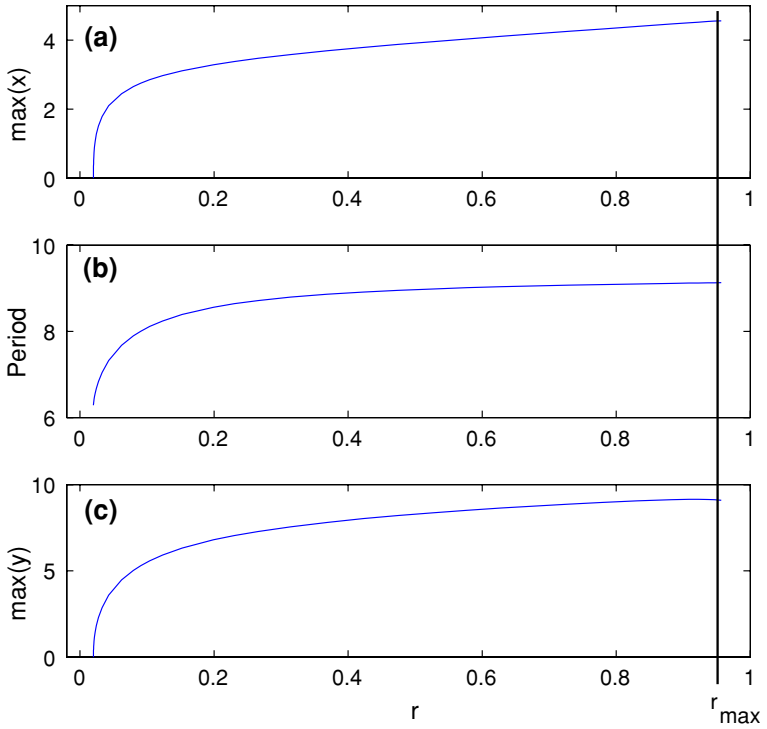

Fig. 3 The maximum of the **a** susceptible $\max(x)$ and **c** infectious $\max(y)$ as well as the **b** period as a function of feedback from the fraction of individuals who become resusceptible, r. $r_{\max} \approx 0.95$ ($a = 1.41$, $b = 0.71$, $\epsilon = 0.014$, and $\tau = 3\pi/2$)

Sect. 5, given by

$$B^2 = \pm\frac{1}{\omega c}(r - r_h), \tag{25}$$

where ω, the constant c and the choice of sign depend upon the delay time τ. Thus, the time duration that individuals are temporarily immune again determines whether the bifurcation is supercritical or subcritical.

For $\tau < \pi$ the branch of bifurcations is always supercritical, and as the delay is decreased the value of r_h increases. Thus, shorter delay times require a larger resusceptible fraction to initiate oscillations. As described in the previous section, for $\tau < \tau_{\min}$ the endemic state is stable for all physically valid values of r and no oscillations will occur. Finally, we note that for the shorter values of delay that occur for $\tau < 3\pi/2$, there is not a secondary bifurcation of a delay mode for $r < r_{\max}$. Thus, the oscillations that initially bifurcate remain stable throughout the range of r, and there are no other stable solutions.

In Fig. 5b we examine values of the delay to the right of the minimum at $\tau = 3\pi/2$ but less than $5\pi/2$. When $\tau = 5.2$ the Hopf bifurcation point is given by Eq. 18, the dashed curve is the analytical prediction given by Eq. 24. When $\tau = 6.7$ the

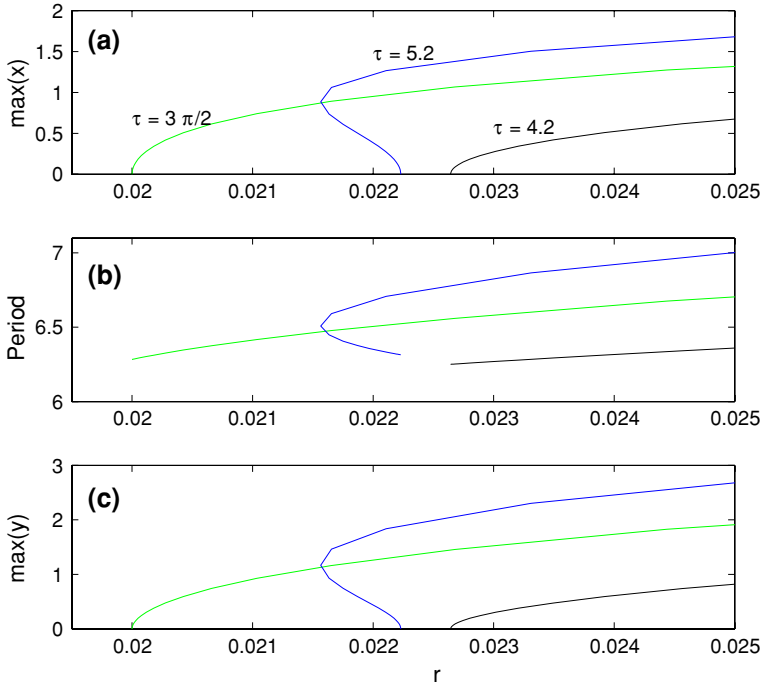

Fig. 4 The maximum values of the susceptible $\max(x)$ and infectious population $\max(y)$ as well as the period local to the Hopf bifurcation point r_h for different values of the delay τ ($a = 1.41$, $b = 0.71$, and $\epsilon = 0.014$)

bifurcation is that of the delay mode with $m = 1$ with bifurcation point given by Eq. 20, the dashed curve is our analytical prediction given by Eq. 25. For these values of the delay the bifurcation is always subcritical and the solutions are stable on the upper part of the branch of periodic solutions. Thus, there is bistability in the interval from the value of r at the left limit point to the Hopf bifurcation point at r_h; depending upon the initial conditions, the disease may be at the endemic steady state or experiencing sizable epidemics.

In Fig. 5c we examine the transition of the primary bifurcation from the $m = 2$ delay mode to the $m = 3$ delay mode, which occurs for $\tau \approx 5\pi/2$. The transition occurs for the value of τ at the intersection $m = 2$ and $m = 3$ neutral stability curves in Fig. 1. $\tau = 7.2$ and 7.9 correspond to the $m = 2$ delay mode so that the primary bifurcation is subcritical. $\tau = 8.15$ and 8.5 correspond to the $m = 3$ delay mode and are supercritical. In Sect. 5 we show that delay modes with m even are subcritical and delay modes with m odd are supercritical.

As the delay is increased, the general pattern described in the previous paragraphs repeats. Natural modes corresponding to delays that are left of the minimum of the neutral stability curve are supercritical, while to the right they are subcritical. As τ is varied through the minimum, there is a continuous deformation of the bifurcation curve from supercritical to subcritical. On the other hand, near the intersection of the

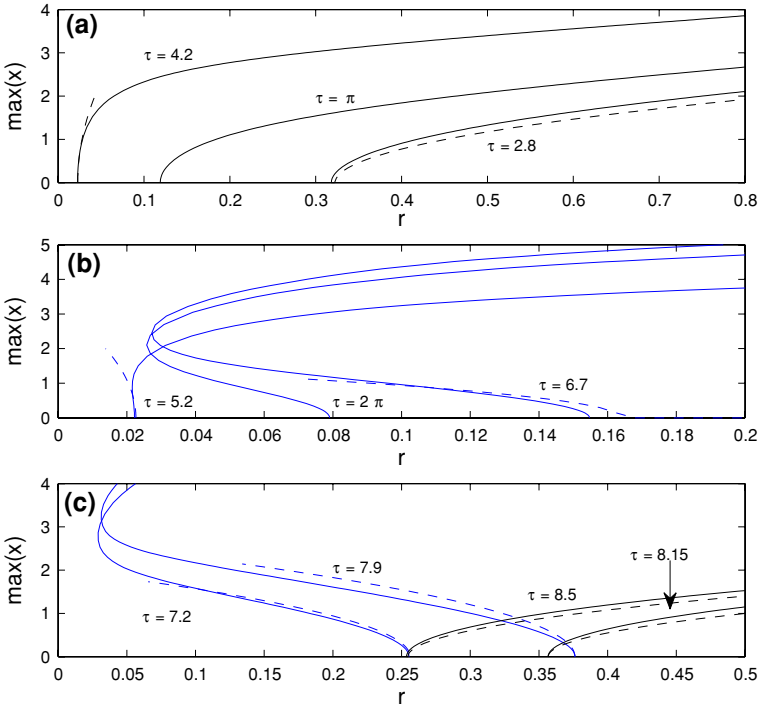

Fig. 5 The maximum value of the susceptible population for different values of the delay ($a = 1.41$, $b = 0.71$, and $\epsilon = 0.014$). The $\tau = \pi$ is unable to be described by our bifurcation results for either the natural or delay modes. *Solid curves* correspond to the numerically computed bifurcation diagrams, while the *dashed solid* were computed numerically. There are no analytical predictions for $\tau = n\pi$ because these values represent singular points of the analysis

branches of the neutral stability curve for the delay modes, the primary bifurcation discontinuously switches from subcritical to supercritical; this is because it is not a continuous deformation of the same branch of solutions, but instead corresponds to switching from the m even to the m odd branch. However, when the supercritical bifurcation for m odd becomes primary, it is followed for only a slightly larger value of r by the subcritical branch for m even, whose upper branch still describes stable oscillations. In this case there can be multi-stability between the small amplitude oscillations of the m-odd oscillations and the larger amplitude oscillations of the m-even oscillations. This multi-stability was described in detail by Pieroux et al. (1994).

3.2.4 Biological mechanisms

In Fig. 6 we consider larger values of the resusceptible parameter r when the oscillations are pulsating solutions. The dashed curves for the amplitude are given by

$$x_f = \frac{2\tau - d_1(1-r)}{2 + d_2(1-r)} - \frac{\epsilon\tau^2}{6r}\left[(2+r)(a-b) + 2(1-r)b\right], \tag{26}$$

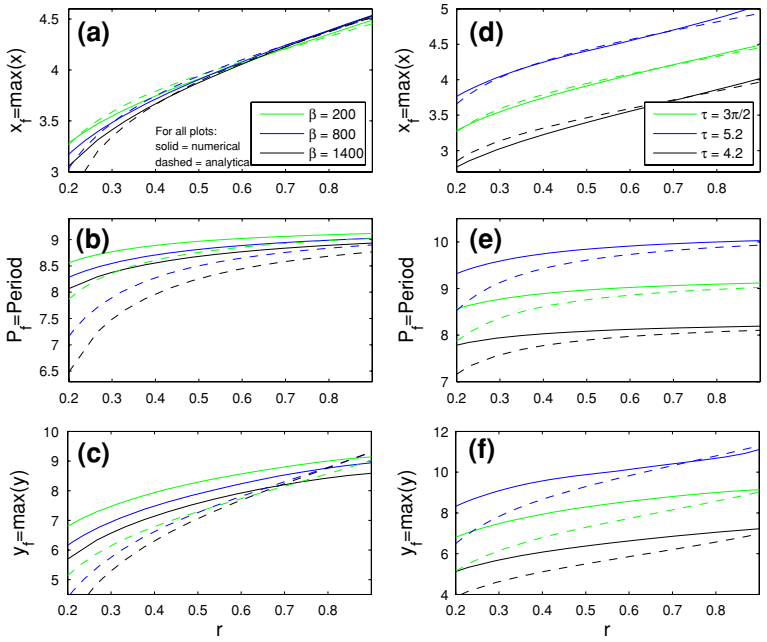

Fig. 6 **a–c** The maximum of the susceptibles x_f, the period P_f, and the maximum of the infectious y_f, as a function of the resusceptible fraction for several values of the transmission coefficient β and $\tau = 3\pi/2$. $\beta = 200$ corresponds to $\epsilon = 0.014$, $a = 1.41$ and $b = 0.71$, $\beta = 800$ to $\epsilon = 0.028$, $a = 1.07$ and $b = 0.94$, and $\beta = 1{,}400$ to $\epsilon = 0.037$, $a = 1.04$ and $b = 0.96$. **d–f** x_f, the period P_f, and y_f for several values of the delay τ with $\epsilon = 0.014$, $a = 1.41$ and $b = 0.71$. *Solid curves* correspond to the analytical predictions of the map, while the *dashed curves* were computed numerically

where x_f represents the amplitude of the oscillation of the susceptible class (the constants d_j and results for the period and the the oscillation amplitudes of the infectious class are all given in Sect. 6). Note that to leading order the delay time and the resusceptible fraction have the strongest affect on determining the amplitude (and period) of the pulsating epidemics. The effect of the biological parameters via ϵ, a and b are a smaller effect.

In Fig. 6a–c we show the effect of changing β. We see that as the transmission rate β is increased, both the severity of the epidemics and the period between epidemics decrease. Increasing β increases both of the dissipation coefficients a and b. With greater dissipation, the system exhibits smaller oscillations. Epidemiologically, this can be understood as follows. From Eq. 6 we see that the basic reproductive number $\mathcal{R}_0$ is directly proportional to β such that as β increases, so will $\mathcal{R}_0$. As $\mathcal{R}_0$ increases, the steady-state number of susceptible individuals decreases because more individuals are in the infectious class. With fewer individuals available to become ill, the epidemic spike will be smaller. If there is a smaller reduction in the susceptible individuals, then this allows that compartment to regenerate faster, leading to a subsequent epidemic spike occurring sooner.

In Fig. 6d–f we show the effect of changing the delay τ and the phenomena of period or frequency locking. This refers to when the period is fixed to some multiple

or fraction of the delay and is a generic property of delay systems. Large increases or decreases in the delay will change the integer relationship between the delay and the period. In the righthand panels of Fig. 6, the system is locked to approximately twice the delay time. Slightly increasing or decreasing the delay time leads to an increase or decrease in the period of the pulsating epidemics. For the system to support a longer period of time before the next epidemic, the supply of susceptible individuals must be depleted to a greater degree in the previous epidemic. This in turn requires larger epidemics. Thus, longer delays lead to longer periods via locking and hence larger amplitudes. Similarly, reducing the delay reduces the amplitude of the epidemics.

In the next three sections we present the mathematical analysis used to derive Eqs. 24–26. The bifurcation equations for the small-amplitude oscillations of the internal and external modes when $r \approx r_h$ are derived using the method of multiple scales, modified to account for the delay. The pulsating oscillations that occur when $r = O(1)$ are described by deriving a map that describes the amplitude and period from one pulse to the next. Readers who are not interested in the mathematical analysis can proceed to the final discussion section, Sect. 7, where we summarize and discuss the results of the paper.

4 Small-amplitude oscillations of the natural modes

Just after the Hopf bifurcation for $r > r_h = O(\epsilon)$, the periodic outbreaks have small amplitude and are nearly harmonic. Typically, a weakly nonlinear perturbation method such as multiple scales (Kevorkian and Cole 1996) is used to analyze oscillations local to a Hopf bifurcation point. In the next section, we will do just that to analyze the delay modes. However, in the case of the natural modes, the leading-order bifurcation equation turns out to be "vertical" in that we do not obtain a relationship between the amplitude of the oscillations and the bifurcation parameter r. Pieroux et al. (1994) have resolved this difficulty by instead looking for $O(1)$ amplitude solutions and finding an equation for the slow evolution of the energy to the $\epsilon = 0$ system. Their final result, given by their Eq. A9, which when written in terms of the parameters of our problem, is

$$r = -\frac{\epsilon a}{\sin\tau}\left[1 + B^2\left(\frac{1}{6}\tau\cot\tau + \frac{5}{18} - \frac{4}{9}\cos\tau\right)\right], \tag{27}$$

where $B \approx \max(x)/2$. Note that for $B = 0$ we recover the linear stability result $r = r_h$ of Eq. 18. As discussed in the linear-stability analysis, r must be non-negative such that $\sin\tau < 0$, and the result is valid for delay times in the intervals $\tau \in (\pi, 2\pi), (3\pi, 4\pi), \ldots$.

In Fig. 5a and b we compare the prediction of Eq. 27 with numerical simulations for the case of $\tau = 4.2$ and $\tau = 5.2$, respectively; as first demonstrated by Pieroux et al. (1994), the fit is quite good. When $\tau = 4.2$ the bifurcation is supercritical and the coefficient of B^2 in Eq. 27 is positive. On the other hand, when $\tau = 5.2$ the bifurcation is subcritical and the coefficient of B^2 is negative. The critical value of τ that separates super from subcritical bifurcations is when the coefficient of $B^2 = 0$, which is approximately the minimum of the neutral stability curve when $\tau = n\pi/2, n = 3, 7, 11, \ldots$.

5 Small-amplitude oscillations of the delay modes

In this section, we will use the method of multiple scales (Kevorkian and Cole 1996), modified to take into account the delay term, to describe the delay modes that emerge via Hopf bifurcations when $r = O(1)$, as given by Eq. 20. The calculation is similar to one that we used when analyzing two coupled lasers in Carr et al. (2006). Without delay, the oscillations decay on an $O(\epsilon)$ time scale, which motivates us to introduce the slow time $T = \epsilon t$; time derivatives then become $\frac{d}{dt} = \frac{\partial}{\partial t} + \epsilon \frac{\partial}{\partial T}$. We analyze the nonlinear problem using perturbation expansions in powers of $\epsilon^{1/2}$, e.g., $x(t) = \epsilon^{1/2} x_1(t, T) + \epsilon x_2(t, T) + \cdots$, while the bifurcation parameter is expanded as $r = r_0 + \epsilon r_1 + \cdots$.

We must also consider the effect of the two-time scale assumption on the delay term. With the additional slow time the delay term becomes

$$y(t - \tau) \to y(t - \tau, T - \epsilon\tau). \tag{28}$$

If $r \ll 1$ then the delay term is small and not part of the leading-order problem. In this case, its effect will be recovered at a higher order as part of a solvability condition for the slowly varying amplitude (Pieroux et al. 2000). The multiple-scale analysis for the delay modes is more complicated because $r = O(1)$, which results in a leading-order problem that contains the delay terms. To make analytical progress we need to remove the slow delay from the leading-order problem (Pieroux et al. 2000). Specifically, we assume that $\epsilon\tau \ll 1$ such that the slow argument can be expanded as

$$y(t - \tau) = y(t - \tau, T) - \epsilon\tau \frac{\partial}{\partial T} y(t - \tau, T) + \cdots . \tag{29}$$

The leading-order problem will still contain the delay on the fast time, but the effect of the delay on the slow time is postponed to higher order. The restriction that $\epsilon\tau \ll 1$ implies that our results are applicable when $\tau = o(1/\epsilon)$. Thus, we find that our results fit well when $\tau = O(1)$ but become less accurate for longer delays.

It should be noted that care must be taken when using a series expansion of a delay term in a differential equation. The Taylor series may itself be justified, but using the series expansion can change the stability of limit sets of the differential equation. A simple example is given in Driver (1977), while (Driver et al. 1973; El'sgol'ts and Norkin 1973) provide more theoretical discussions concerning restrictions on the size of the delay. In our presentation we will check the validity of our approximations by comparing our analytical and numerical results.

5.1 Leading order

The leading order $O(\epsilon^{1/2})$ problem is

$$\frac{\partial}{\partial t} X_1(t, T) = L \cdot X_1(t, T) + D \cdot X_1(t - \tau, T), \tag{30}$$

where

$$X_1(t,T) = \begin{pmatrix} x_1(t,T) \\ y_1(t,T) \end{pmatrix}, \quad L = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad D = \begin{pmatrix} 0 & r_0 \\ 0 & 0 \end{pmatrix}. \tag{31}$$

We look for oscillatory solutions of the form $X_1(t,T) = U_1 A(T)\exp(i\omega t)+$c.c., (c.c. represents the complex conjugate) where $A(T)$ is a slowly varying scalar amplitude [to be determined from a solvability condition at $O(\epsilon^{3/2})$]. To find ω and the vector U_1 we substitute our ansatz into Eq. 30 to obtain

$$0 = J \cdot U_1 \quad \text{where} \quad J = \begin{pmatrix} -i\omega & -1 + r_0 e^{-i\omega\tau} \\ 1 & -i\omega \end{pmatrix}. \tag{32}$$

For a non zero solution U_1, we require $\det J = 0$. This results in the same condition obtained from the leading-order linear-stability problem, and we find that

$$\omega = \frac{m\pi}{\tau}, \quad m = \text{an integer}. \tag{33}$$

and

$$r_0 = \pm(1 - \omega^2), \tag{34}$$

where, as in Sect. 3.1, the positive solution is taken if m is even and the negative solution is taken if m is odd. Finally, we find that

$$U_1 = \begin{pmatrix} i\omega \\ 1 \end{pmatrix}. \tag{35}$$

5.2 Second order

At $O(\epsilon)$ the problem is

$$\frac{\partial}{\partial t} X_2(t,T) = L \cdot X_2(t,T) + D \cdot X_2(t-\tau,T) + F_2, \quad \text{where } F_2 = \begin{pmatrix} 0 \\ x_1 y_1 \end{pmatrix} \tag{36}$$

Because the homogeneous problem is the same as the $O(\epsilon^{1/2})$, problem we can, without loss of generality, set the homogeneous solution to 0. The inhomogeneous term F_2 is proportional to $\exp(i2\omega t)$ so that the solution is

$$X_2(t,T) = A(T)^2 U_2 e^{i2\omega t} + \text{c.c.}, \tag{37}$$

where

$$U_2 = \frac{1}{(1-4\omega^2) - r_0} \begin{pmatrix} i\omega(r_0 - 1) \\ -2\omega^2 \end{pmatrix}. \tag{38}$$

5.3 Third order

At $O(\epsilon^{3/2})$ we find the solvability condition that determines the slow-evolution equation for $B(T)$. The $O(\epsilon^{3/2})$ problem is

$$\frac{\partial}{\partial t}X_3(t,T) = L \cdot X_3(t,T) + D \cdot X_3(t-\tau,T) + F_3, \tag{39}$$

where

$$F_3 = \begin{pmatrix} -ax_1 + r_0 y_3(t-\tau,T) - r_0\tau \frac{\partial}{\partial T} y_1(t-\tau,T) - \frac{\partial}{\partial T}x_1 \\ x_1 y_2 + x_2 y_1 - \frac{\partial}{\partial T} y_1 \end{pmatrix}. \tag{40}$$

The vector F_3 has terms proportional to $\exp(i\omega t)$ and $\exp(i2\omega t)$, and the former will lead to solutions of the form $(U_3 + V_3 t)\exp(i\omega t)$. The secular term $V_3 t$ must be eliminated to prevent unbounded solutions for large t, which implies that a solvability condition must be imposed on F_3. The solvability condition is formulated as follows: We look for a solution to Eq. 39 of the form $X_3 = U\exp(i\omega t)$ and at the same time identify the terms in F_3 proportional to $\exp(i\omega t)$. We then obtain an algebraic system of equations for the vector U as

$$0 = J \cdot U + F, \tag{41}$$

where

$$F = \begin{pmatrix} (-i\omega a + r_1 e^{-i\omega\tau})A - (i\omega + r_0\tau e^{-i\omega\tau})\frac{\partial A}{\partial T} \\ ic|A|^2 A - \frac{\partial A}{\partial T} \end{pmatrix}, \tag{42}$$

and

$$c = \begin{cases} \omega & m \text{ even} \\ \frac{\omega(\omega^2+2)}{5\omega^2-2} & m \text{ odd} \end{cases}. \tag{43}$$

For U to have a non-zero solution, the Fredholm alternative requires that $V^H \cdot F = 0$, where V is the solution to $J^H \cdot V = 0$ (the superscript H refers to Hermitian). We find that $V^H = (1, i\omega)$, and the resulting condition for the amplitude $A(T)$ is

$$(i2\omega + r_0\tau e^{-i\omega\tau})\frac{\partial A}{\partial T} = (-i\omega a + r_1 e^{-i\omega\tau})A + i\omega c A|A|^2, \tag{44}$$

5.4 Bifurcation equation

To analyze the solvability condition given by Eq. 44, we let $A(T) = B(T)e^{i\theta(T)}$. The bifurcation equation is determined by considering steady-state solutions to the

equation for B, and we find that

$$B^2 = \pm\frac{1}{\omega c}\left(r_1 - \frac{2\omega^2 a}{r_0\tau}\right) \tag{45}$$

where c is positive and the positive solution is taken if m is odd and the negative solution is taken if m is even. Notice that the second term in the parentheses is the correction to the Hopf bifurcation point as given in Eq. 20. Thus, we have that

$$B^2 = \pm\frac{1}{\omega c}(r_1 - r_{1h}) \tag{46}$$

and the signs indicate that the bifurcation is supercritical if m is odd and subcritical if m is even.

We have used Eq. 44 to generate the dashed curves in Fig. 5a for $\tau = 2.8$, in Fig. 5b for $\tau = 6.7$, and all the dashed curves shown in Fig. 5c; we obtain excellent results for a wide range of values of the delay.

6 Pulsating outbreaks for high resusceptibility

In this section we describe the pulsating solutions that occur for $r = O(1)$. During the time interval from one pulse to the next, there are times when the terms $y(t)$ and $y(t-\tau)$ in Eq. 9 are either large or approximately -1 (see Fig. 7). We will use these observations to find approximations to Eq. 9 that are easier to analyze. Similar to the method of matched asymptotics (Kevorkian and Cole 1996), we solve approximations of Eq. 9 on separate subintervals, defined by the relative scaling of $y(t)$ and $y(t-\tau)$. Specifically, we mark the beginning of a pulse where $y(t_0) = 0$ and the end of the pulse where $y(t_0') = 0$. Over the short time interval (t_0, t_0') the infected population $y(t)$ is large, while $y(t-\tau)$ is approximately -1 and the susceptible population rapidly decreases to its minimum. Following the pulse during a longer subinterval $(t_0', t_0+\tau)$, both $y(t)$ and $y(t-\tau)$ are approximately -1, while the susceptible population increases slowly. During the delayed pulse $y(t) \approx -1$ and $y(t-\tau)$ is large. Finally, there is another long subinterval following the delayed pulse where both $y(t)$ and $y(t-\tau)$ are approximately -1 and the susceptible population $x(t)$ increases slowly to its maximum. The initial condition for each subinterval comes from the terminal condition of the previous subinterval. The end result is a map describing the time and amplitude of the next pulse, t_1, $x(t_1)$ and $y(t_1)$ in terms of the present time and amplitude, t_0, $x(t_0)$ and $y(t_0)$. This technique has been used to analyze the pulsating output of lasers (Schwartz and Erneux 1994; Carr 2003; Carr et al. 2000). Readers not interested in these details may proceed directly to the analysis of the resulting map in Sect. 6.4.

6.1 Subinterval approximations

We now show the specific approximations and solutions in each subinterval. We first consider the time intervals when the delay pulse is small or, $y(t-\tau) \approx -1$. With

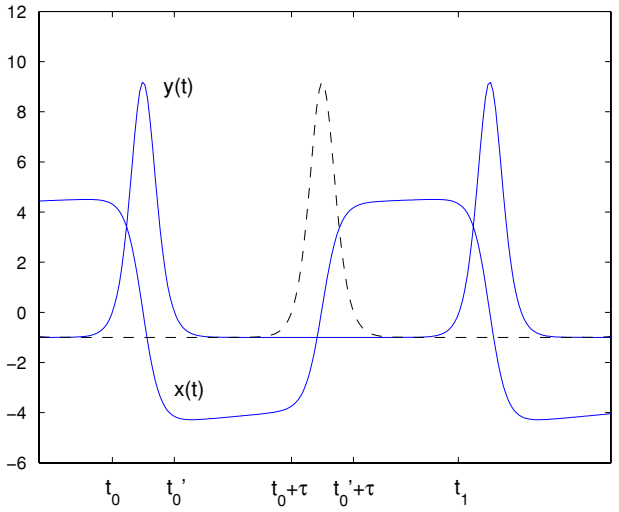

Fig. 7 The susceptible, $x(t)$, and infected, $y(t)$, populations for $r = 0.9$. The time-delayed pulse is shown as a *dashed line* ($a = 1.41$, $b = 0.71$, $\epsilon = 0.014$, and $\tau = 3\pi/2$)

reference to Fig. 7, this is everywhere *outside* the interval $t \in [t_0 + \tau, t_0' + \tau]$. Using the change of variable $y + r = Y$, Eqs. 9 become

$$\begin{aligned} \frac{dx}{dt} &= -Y - \epsilon x(A + bY), \\ \frac{dY}{dt} &= x(\rho + Y), \end{aligned} \tag{47}$$

where $A = a - br$ and $\rho = 1 - r$. We now use Eqs. 47 to describe the system for times when the delayed pulse is small ($y(t - \tau) \approx -1$).

Pulse when $t \in [t_0, t_0']$ Define the times $t = t_0$ and $t = t_0' = t_0 + \Delta$ as the beginning and end of the present pulse, where Δ refers to the pulse width. More specifically, the start of the pulse occurs when x is a maximum, dx/dt at t_0 is zero. Similarly, the end of the pulse occurs when x is a minimum.

During the pulse, $Y \gg \rho$ and $Y \gg A/b$ so Eqs. 47 are approximated by

$$\begin{aligned} \frac{dx}{dt} &= -Y - \epsilon bxY, \\ \frac{dY}{dt} &= xY. \end{aligned} \tag{48}$$

These can be solved in the phase plane by determining the equation for $\frac{dY}{dx}$ whose solution is

$$Y(t) = -\frac{1}{\epsilon b}\left\{x(t) - x(t_0) - \frac{1}{\epsilon b}\left[\ln(1 + \epsilon bx(t)) - \ln(1 + \epsilon bx(t_0))\right]\right\}, \tag{49}$$

where $Y(t_0) = 0$ via Eq. 48 because we require $dx/dt = 0$ at $t = t_0$. Similarly, at the end of the pulse $Y(t_0')$ is also zero so that $x(t_0')$ satisfies

$$0 = x(t_0') - x(t_0) - \frac{1}{\epsilon b}\left[\ln(1 + \epsilon b x(t_0')) - \ln(1 + \epsilon b x(t_0))\right] \tag{50}$$

and after expanding for $\epsilon \ll 1$, we obtain

$$x(t_0') = -x(t_0) + \frac{2}{3}\epsilon b x^2(t_0) + O(\epsilon^2) \tag{51}$$

(expand the natural logs and $x(t_0')$ in powers of ϵ). The peak value of the pulse, which corresponds to the peak in the infectious population, occurs when $dY/dt = 0$ and, hence, from Eq. 48 when $x(t_p) = 0$. Thus, from Eq. 49 we have:

$$Y(t_p) = \frac{1}{\epsilon b}\left[x(t_0) - \frac{1}{\epsilon b}\ln(1 + \epsilon b x(t_0))\right]. \tag{52}$$

First "outer" interval when $t \in [t_0', t_0 + \tau]$. In the next interval from $t = t_0'$ to $t = t_0 + \tau$, $Y(t) \approx -\rho$, and Eq. 47 are approximated by

$$\begin{aligned} \frac{dx}{dt} &= \rho - \epsilon\eta x, \\ \frac{dY}{dt} &= x(\rho + Y), \end{aligned} \tag{53}$$

where $\eta = (a - b)$. The equation for x can be solved first and the result used to find Y; we obtain

$$\begin{aligned} x(t_0 + \tau) &= \left[x(t_0') - \frac{\rho}{\epsilon\eta}\right] e^{-\epsilon\eta(\tau - \Delta)} + \frac{\rho}{\epsilon\eta}, \\ Y(t_0 + \tau) &= -\rho + \left[\rho + Y(t_0')\right] e^{F(t_0 + \tau, t_0')}, \end{aligned} \tag{54}$$

where $\Delta = t_0' - t_0$ and

$$\begin{aligned} F(t, t_0') &= x(t_0')(t - t_0') + \frac{1}{2}\rho(t - t_0')^2 \\ &\quad - \frac{1}{2}\epsilon\eta\left[x(t_0')(t - t_0')^2 + \frac{1}{3}\rho(t - t_0')^3\right], \\ F(t_0 + \tau, t_0') &= x(t_0')T_1 + \frac{1}{2}\rho T_1^2 - \frac{1}{2}\epsilon\eta\left[x(t_0')T_1^2 + \frac{1}{3}\rho T_1^3\right], \\ T_1 &= \tau - \Delta. \end{aligned} \tag{55}$$

T_1 represents the duration of time in the first outer subinterval.

Delayed pulse when $t \in [t_0 + \tau, t_0' + \tau]$. The delayed pulse occurs in the interval from $t = t_0 + \tau$ to $t = t_0' + \tau$. During this interval we consider the model in original variables with y instead of Y. $y(t - \tau) \gg 1$ and $y \approx -1$ and Eq. 9 are approximated by

$$\begin{aligned} \frac{dx}{dt} &= 1 - \epsilon\eta x + ry(t - \tau), \\ \frac{dy}{dt} &= 0. \end{aligned} \tag{56}$$

To leading order y is unaffected by the delay term so that

$$y(t_0' + \tau) = y(t_0 + \tau). \tag{57}$$

The equation for x is linear with $y(t - \tau)$ serving as a known forcing term and, hence, can be solved with an integrating factor to obtain

$$\begin{aligned} x(t_0' + \tau) = x(t_0 + \tau)e^{-\epsilon\eta\Delta} + \frac{1}{\epsilon\eta}(1 - e^{-\epsilon\eta\Delta}) \\ + re^{-\epsilon\eta(t_0' + \tau)} \int_{t_0 + \tau}^{t_0' + \tau} e^{\epsilon\eta s} y(s - \tau) ds, \end{aligned} \tag{58}$$

where $\Delta = t_0' - t_0$ is the width of the pulse. The effect of the delay term is to cause a jump in $x(t)$ proportional to the area of the original pulse in $y(t)$. What remains is to evaluate the integral.

With the change of variable $s - \tau \to s$ we analyze

$$I_1 = re^{-\epsilon\eta t_0'} \int_{t_0}^{t_0'} e^{\epsilon\eta s} y(s) ds, \tag{59}$$

which requires the solution for y in the time interval of the original pulse. The latter was described in terms of the variable $Y = y + r$ so that the integral becomes

$$I_1 = -r^2 e^{-\epsilon\eta t_0'} \int_{t_0}^{t_0'} e^{\epsilon\eta s} ds + re^{-\epsilon\eta t_0'} \int_{t_0}^{t_0'} e^{\epsilon\eta s} Y(s) ds. \tag{60}$$

The first integral in I_1 can be evaluated directly. For the second integral we first note from Eqs. 48 that

$$Y = \frac{-\frac{dx}{dt}}{1 + \epsilon bx} \approx -\frac{dx}{dt} + \epsilon bx \frac{dx}{dt} + O(\epsilon^2), \tag{61}$$

to give

$$I_1 = -r^2 \frac{1}{\epsilon\eta}(1 - e^{-\epsilon\eta\Delta}) + re^{-\epsilon\eta t_0'} \int_{t_0}^{t_0'} e^{\epsilon\eta s} \left(-\frac{dx}{dt} + \epsilon b x \frac{dx}{dt} + O(\epsilon^2) \right) ds. \quad (62)$$

Substituting the result for I_1 into Eq. 58 for x at the end of delayed pulse we have

$$x(t_0' + \tau) = x(t_0 + \tau)e^{-\epsilon\eta\Delta} + \frac{1}{\epsilon\eta}(1 - r^2)(1 - e^{-\epsilon\eta\Delta}) + re^{-\epsilon\eta t_0'} \int_{t_0}^{t_0'} e^{\epsilon\eta s} \left(-\frac{dx}{dt} + \epsilon b x \frac{dx}{dt} + O(\epsilon^2) \right) ds, \quad (63)$$

We must now evaluate the integral

$$I_2 = re^{-\epsilon\eta t_0'} \int_{t_0}^{t_0'} e^{\epsilon\eta s} \left(-\frac{dx}{dt} + \epsilon b x \frac{dx}{dt} + O(\epsilon^2) \right) ds,$$
$$= re^{-\epsilon\eta t_0'} \int_{t_0}^{t_0'} e^{\epsilon\eta s} \left(-\frac{dx}{dt} + \epsilon b \frac{1}{2} \frac{d}{dt}(x^2) + O(\epsilon^2) \right) ds. \quad (64)$$

Each of the integrals in I_2 can be evaluated *by parts* to give

$$I_2 = r \left[-x(t_0') + e^{-\epsilon\eta\Delta} x(t_0) + \frac{1}{2}\epsilon b \left(x(t_0')^2 - e^{-\epsilon\eta\Delta} x(t_0)^2 \right) \right] + re^{-\epsilon\eta t_0'} \left[\epsilon\eta \int_{t_0}^{t_0'} e^{\epsilon\eta s} x ds - \epsilon^2 b\eta \int_{t_0}^{t_0'} e^{\epsilon\eta s} x^2 ds + O(\epsilon^3) \right]. \quad (65)$$

The last integral with x^2 is explicitly $O(\epsilon^2)$ and will be ignored. It turns out that the contribution of the first integral is also $O(\epsilon^2)$. The exponential in the integrand can be expanded as the following:

$$e^{\epsilon\eta s} \approx [1 + \epsilon\eta s + O(\epsilon^2)]x(s) \approx x(s) + O(\epsilon). \quad (66)$$

The integral of the first term $x(s)$ over the interval of the pulse is zero because $x(s)$ is odd to leading order (see Fig. 7 and Eq. 51). Thus, the first integral is $O(\epsilon)$ and because it is multiplied by ϵ, its contribution is $O(\epsilon^2)$. The result is that

$$I_2 = r \left[-x(t_0') + e^{-\epsilon\eta\Delta} x(t_0) + \frac{1}{2}\epsilon b \left(x(t_0')^2 - e^{-\epsilon\eta\Delta} x(t_0)^2 \right) \right] + O(\epsilon^2). \quad (67)$$

Finally, we use Eq. 51 to substitute for $x(t_0')$ to give

$$I_2 = r\left[x(t_0)\left(1+e^{-\epsilon\eta\Delta}\right) - \frac{2}{3}\epsilon b x(t_0)^2 + \frac{1}{2}\epsilon b x(t_0)^2\left(1-e^{-\epsilon\eta\Delta}\right)\right] + O(\epsilon^2). \quad (68)$$

Thus, Eq. 63 for x at the end of the delayed pulse is

$$\begin{aligned} x(t_0'+\tau) = {} & x(t_0+\tau)e^{-\epsilon\eta\Delta} + \frac{1}{\epsilon\eta}(1-r^2)(1-e^{-\epsilon\eta\Delta}) \\ & + r\left[x(t_0)\left(1+e^{-\epsilon\eta\Delta}\right) - \frac{2}{3}\epsilon b x(t_0)^2 + \frac{1}{2}\epsilon b x(t_0)^2\left(1-e^{-\epsilon\eta\Delta}\right)\right] \\ & + O(\epsilon^2). \end{aligned} \quad (69)$$

The last step to determining a reasonably simple equation for $x(t_0'+\tau)$ is to expand the exponential functions for $\epsilon\eta\Delta = O(\epsilon)$. Doing so we obtain

$$\begin{aligned} x(t_0'+\tau) = {} & x(t_0+\tau)(1-\epsilon\eta\Delta) + (1-r^2)\Delta\left(1-\frac{1}{2}\epsilon\eta\Delta\right) \\ & + r\left[x(t_0)(2-\epsilon\eta\Delta) - \frac{2}{3}\epsilon b x(t_0)^2\right] + O(\epsilon^2). \end{aligned} \quad (70)$$

Second "outer" interval when $t \in [t_0'+\tau, t_1]$. In the last time interval, both $Y(t)$ and $Y(t-\tau)$ are again approximately $-\rho$. The time t_1 is defined when x again reaches a maximum, which we can identify as occurring when $Y(t_1) = 0$. We solve the same equations as for $t \in [t_0', t_0+\tau]$, and find that

$$\begin{aligned} x(t_1) &= \left[x(t_0'+\tau) - \frac{\rho}{\epsilon\eta}\right]e^{-\epsilon\eta(P-(\tau+\Delta))} + \frac{\rho}{\epsilon\eta}, \\ Y(t_1) &= -\rho + \left[\rho + Y(t_0'+\tau)\right]e^{F(t_1,t_0'+\tau)}, \end{aligned} \quad (71)$$

where $P = t_1 - t_0$ is the total time between the initiation of two pulses. $F(t_1, t_0'+\tau)$ is defined similarly to Eq. 55 so that here we have

$$\begin{aligned} F(t_1, t_0'+\tau) &= x(t_0'+\tau)T_2 + \frac{1}{2}\rho T_2^2 - \frac{1}{2}\epsilon\eta\left[x(t_0'+\tau)T_2^2 + \frac{1}{3}\rho T_2^3\right], \\ T_2 &= P - (\tau+\Delta), \end{aligned} \quad (72)$$

and T_2 is the total time in the second outer subinterval.

6.2 Pulse width Δ

The pulse width $\Delta = t_0' - t_0$ is defined to be the short interval of time between the extrema of $x(t)$. Unfortunately, Δ is neither small nor constant and can not be ignored.

To find an approximation for the pulse width, we consider the system when the delay term is at its minimum with $y(t-\tau) \approx -1$. For the purposes of finding the pulse width, we will also consider the damping terms to be negligible. Thus, we consider the conservative system

$$\begin{aligned} \frac{dx}{dt} &= -Y, \\ \frac{dY}{dt} &= x(\rho + Y), \end{aligned} \tag{73}$$

To find Δ we need an explicit solution for $x(t)$ in order to learn when the maximum and minimum occur. Unfortunately, while Eqs. 73 has the first integral

$$C = \frac{1}{2}x^2 + Y - \rho \ln(\rho + Y), \tag{74}$$

there is no exact solution for $x(t)$. However, Pieroux and Erneux (1996) have constructed a uniform asymptotic expansion for $x(t)$ in the case when $\rho = 1$. We have followed their analysis to derive a uniform solution for $x(t)$ for general ρ, and it is given by

$$x(t) \approx \rho t - \frac{e^{\xi}}{\frac{\rho t_c}{C_0} + \frac{1}{\rho t_c} e^{\xi}}, \qquad \xi = t_c \rho (t - t_c). \tag{75}$$

$$t_c \approx \frac{\sqrt{2C_0}}{\rho}\left(1 + \frac{1}{2}\frac{\rho}{C_0}\ln C_0\right) \tag{76}$$

$$C_0 = \frac{1}{2}x(t_0)^2 - \rho \ln \rho, \qquad Y(t_0) = 0. \tag{77}$$

Given $x(t)$, we find when $dx/dt = 0$ and used those times to determine the pulse width to be

$$\Delta \approx \sqrt{\frac{2}{C_0}} \ln\left(\frac{4C_0}{\rho}\right)\left(1 - \frac{1}{2}\frac{\rho}{C_0}\ln C_0\right). \tag{78}$$

In practice, we find that we can ignore the correction terms to both C_0 and Δ. Specifically, we let $C_0 = (1/2)x(t_0)^2$ in the equation for Δ and obtain

$$\Delta \approx \frac{4}{x(t_0)} \ln\left(\sqrt{\frac{2}{\rho}} x(t_0)\right) \tag{79}$$

In Fig. 8a we compare Eq. 79 to results from numerical simulations of Eq. 73. The thin solid curve is the pulse width based on numerical simulation of the ordinary differential equations, Eqs. 73. The thick solid curve is the pulse width of the DDEs, Eqs. 9. Our first observation is that the pulse width of the ODE model is a good approximation of the pulse width of the DDE model. The dashed curve is the pulse width based on

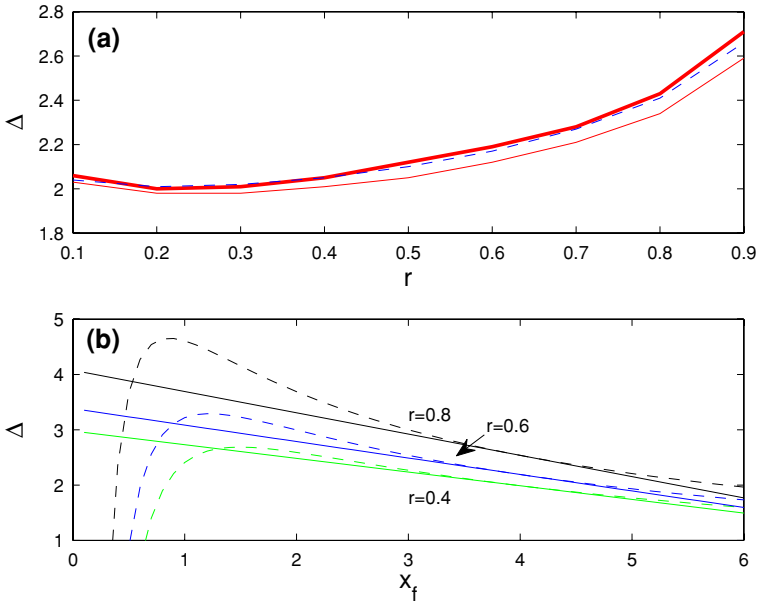

Fig. 8 The pulse width as a function of the feedback rate r. **a** The *thin solid curve* is the numerically computed pulse width for the system without delay, while the *dashed curve* is the analytical approximation of the pulse width, Eq. 79, for the system without delay. The *thick solid curve* is the pulse width of the delayed system. **b** For fixed values of r, the linear approximation of the pulse width given by Eq. 80 compared to Eq. 79

Eq. 79, where we used the numerical value of $x(t_0)$ to find Δ. We see that Eq. 79 does an excellent job of describing the pulse width of the delay system over the full range of the feedback parameter r.

Unfortunately, using Eq. 79 for Δ, which is given in terms of a natural log function, makes it impossible to determine explicit final answers for x and the period. Thus, we use a linear approximation for Δ that we find does a very good job of fitting the actual function and is shown in Fig. 8b. We find that

$$\Delta = d_1 + d_2 x(t_0), \tag{80}$$

$$d_1 = \frac{4}{x_c}\left(2\ln\left(\sqrt{\frac{2}{\rho}}x_c\right) - 1\right), \quad d_2 = \frac{4}{x_c^2}\left(1 - \ln\left(\sqrt{\frac{2}{\rho}}x_c\right)\right), \tag{81}$$

where $x = x_c$ is the expansion point. We have to choose x_c using some measure of best fit. Clearly, the value of $x = x_c$ should be within the range of numerically computed amplitudes. However, we have not tried to optimize our choice beyond the observation criteria that we now describe. In Fig. 9 we have selected $x_c = 4$ and observe that the numerically computed bifurcation curves for $\beta = 200$ and $\beta = 800$ intersect at approximately the same value of r as the corresponding analytically computed curves.

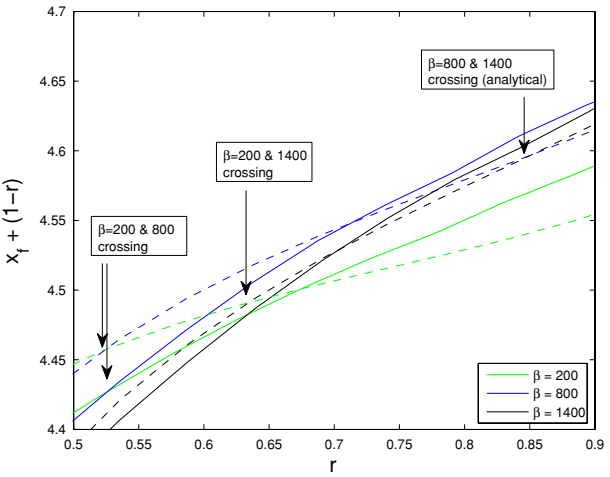

Fig. 9 The maximum of the susceptible population x_f but shifted by $1 - r$; otherwise, same parameter values as in Fig. 6a

The intersections for the $\beta = 200$ and $\beta = 1{,}400$ curves are essentially equivalent. However, the intersections of the $\beta = 800$ and $\beta = 1{,}400$ curves do not line up as well. We matched these intersection points as a way to very coarsely choose between $x_c = 4$ and, for example, $x_c = 5$, but did not try to be more precise than this. That said, in general, our final results are relatively insensitive to the choice of x_c such that we observe significant deviations only for $x_c \approx 10$ or $x_c \approx 1$.

6.3 Constructing the map

The map is determined by patching together the results on each subinterval to determine a relationship for $x(t_1)$ and t_1 in terms of $x(t_0)$ and t_0. More specifically, we have derived the following relationships:

$$\begin{aligned}
&\text{From Eq. 51:} \quad x(t_0') = f_1(x(t_0)),\\
&\text{From Eq. 54:} \quad x(t_0 + \tau) = f_2(x(t_0')),\\
&\text{From Eq. 70:} \quad x(t_0' + \tau) = f_3(x(t_0 + \tau)),\\
&\text{From Eq. 71:} \quad x(t_1) = f_4(x(t_0' + \tau)).
\end{aligned} \tag{82}$$

Thus, in general, if we let $x_n = x(t_0)$ be the current maximum value of x, and x_{n+1} be the next maximum, then the map is given by the composition of the relationships as

$$x_{n+1} = f_4(f_3(f_2(f_1(x_n)))). \tag{83}$$

After making the explicit substitutions and expanding all exponentials for $\epsilon \ll 1$, we find that

$$x_{n+1} = \rho\left[T_3(1-\frac{1}{2}\epsilon\eta T_3) - \epsilon\eta T_1\Delta_n\right] + (1-r^2)\Delta_n\left[(1-\epsilon\eta T_2) - \frac{1}{2}\epsilon\eta\Delta_n\right] + x_n\left[2r(1-\epsilon\eta T_2) - (1-\epsilon\eta T_3) + \epsilon\eta\rho\Delta_n\right] + \frac{2}{3}\epsilon b\rho x_n^2 + O(\epsilon^2), \quad (84)$$

where

$$T_1 = \tau - \Delta_n, \quad T_2 = P_n - (\tau + \Delta_n), \quad \text{and} \quad T_3 = T_1 + T_2. \quad (85)$$

$\Delta_n = \Delta(x_n)$ is the pulse width of the current pulse and is given by Eq. 79, which we reproduce here using the map notation:

$$\Delta_n = d_1 + d_2 x_n \quad (86)$$

Finally, $P_n = t_{n+1} - t_n$ is the total time from start of the current pulse to the next and has yet to be determined.

To determine P_n we use the results for Y. Recall that we define Y to have the same value at the beginning and the end of the pulse, i.e., $Y(t_0') = Y(t_0)$. We also assume that when the delay pulse is large, this has little effect on Y, i.e., $Y(t_0'+\tau) = Y(t_0+\tau)$. Thus, taking the composition of results for Y from Eqs. 71 and 54, we have that

$$Y(t_{n+1}) = -\rho + [\rho + Y(t_n)]\, e^{F(t_0+\tau,t_0')+F(t_1,t_0'+\tau)}, \quad (87)$$

The beginning and end of the pulse are defined when $dx/dt = 0$ and, hence, $Y = 0$. Thus, Eq. 87 requires that $F(t_0+\tau, t_0') + F(t_1, t_0'+\tau) = 0$. Writing this in terms of the map variables, we see that the following condition must be satisfied:

$$\begin{aligned}
&\rho\left[\frac{1}{2}T_3^2 - \frac{1}{6}\epsilon\eta T_3^3 - \epsilon\eta\Delta_n T_1 T_2\right] \\
&+(1-r^2)\Delta_n T_2\left[1 - \frac{1}{2}\epsilon\eta(\Delta_n + T_2)\right] \\
&-x_n\left[T_3 - \frac{1}{2}\epsilon\eta T_3^2 - 2r\left(T_2 - \frac{1}{2}\epsilon\eta T_2^2\right) - \epsilon\eta\rho\Delta_n T_2\right] \\
&+\frac{2}{3}\epsilon b x_n^2(T_3 - rT_2) + O(\epsilon^2) = 0.
\end{aligned} \quad (88)$$

In summary, we have a map from $(x_n, t_n) \mapsto (x_{n+1}, t_{n+1})$ given by

$$\begin{aligned}
&\text{From Eq. 86:} \quad x_n \mapsto \Delta_n \\
&\text{From Eq. 88:} \quad (x_n, \Delta_n) \mapsto P_n \\
&\text{From Eq. 84:} \quad (x_n, \Delta_n, P_n) \mapsto x_{n+1}
\end{aligned} \quad (89)$$

and $t_{n+1} = t_n + P_n$.

6.4 Fixed points of the map

Periodic solutions of the original DDEs correspond to fixed points of the map, which are described by the following coupled set of equations:

$$\begin{aligned} &-x_f + \rho\left[T_3(1-\frac{1}{2}\epsilon\eta T_3) - \epsilon\eta T_1\Delta\right] \\ &+(1-r^2)\Delta\left[(1-\epsilon\eta T_2) - \frac{1}{2}\epsilon\eta\Delta\right] \\ &+x_f\left[2r(1-\epsilon\eta T_2) - (1-\epsilon\eta T_3) + \epsilon\eta\rho\Delta\right] + \frac{2}{3}\epsilon b\rho x_f^2 = 0, \end{aligned} \tag{90}$$

where

$$T_1 = \tau - \Delta, \quad T_2 = P_f - (\tau + \Delta), \quad \text{and} \quad T_3 = T_1 + T_2, \tag{91}$$

$$\Delta = d_1 + d_2 x_f, \tag{92}$$

and

$$\begin{aligned} &\rho\left[\frac{1}{2}T_3^2 - \frac{1}{6}\epsilon\eta T_3^3 - \epsilon\eta\Delta T_1 T_2\right] + (1-r^2)\Delta T_2\left[1 - \frac{1}{2}\epsilon\eta(\Delta + T_2)\right] \\ &-x_f\left[T_3 - \frac{1}{2}\epsilon\eta T_3^2 - 2r(T_2 - \frac{1}{2}\epsilon\eta T_2^2) - \epsilon\eta\rho\Delta T_2\right] \\ &+\frac{2}{3}\epsilon b x_f^2(T_3 - rT_2) = 0. \end{aligned} \tag{93}$$

A general solution for x_f and P_f is not possible without additional approximations. With Δ given by Eq. 91, Eqs. 90 and 93 are algebraic in x and P, and we solve for them using a perturbation expansion. We let $x_f = x_0 + \epsilon x_1 + \cdots$, $P_f = P_0 + \epsilon P_1 \cdots$ and then collect terms by powers of ϵ. At $O(1)$ we find that

$$x_0 = \frac{P_0 - d_1\rho}{2 + d_2\rho}, \qquad P_0 = 2\tau. \tag{94}$$

Thus, to leading order the period is locked to be twice the delay time. If we ignore the effect of the pulse width, then $x_0 \sim P_0/2$, which is consistent with previous map derivations for systems related to Eqs. 9 (Carr et al. 2000; Schwartz and Erneux 1994).

Inclusion of the pulse width in the $O(\epsilon)$ corrections results in quite complicated and analytically intractable solutions for x_1 and P_1. Thus, we have chosen to set $\Delta = 0$ ($d_1 = d_2 = 0$) in the $O(\epsilon)$ problem and still get excellent results when we compare against numerical solutions. With this final simplification, the results for the x and P,

including both the leading order and correction terms, are

$$\begin{aligned} x_f &= \frac{2\tau - d_1(1-r)}{2 + d_2(1-r)} - \frac{\epsilon\tau^2}{6r}\left[(2+r)\eta + 2(1-r)b\right], \\ P_f &= 2\tau - \frac{\epsilon\tau^2}{3r}\left[(2+r)\eta + 2b\right], \end{aligned} \tag{95}$$

where we have made the substitution $P_0 = 2\tau$ into the equation for x.

Finally, we use the value of x with Eq. 52 to determine the peak value of the pulse in y, which corresponds to the maximum deviation from the endemic infectious population equilibrium:

$$y_f = -r + \frac{1}{2}x_f^2 - \frac{1}{3}\epsilon b x_f^3. \tag{96}$$

Thus, Eqs. 95 and 96 describe how the maximum value of the susceptible and infectious populations during epidemics, as well as the time period between the epidemics, depends upon the model parameters and the delay time τ.

6.5 Comparison of analytical and numerical bifurcation results

In Fig. 6d–f we compare the map results from Eqs. 95 and 96 (solid curves) to numerical simulations (dashed curves). Recall that in deriving the map we assumed large and narrow pulsating solutions when $r = O(1)$. As a result, in each plot, agreement between the numerical and analytical results improves as r is increased further away from the Hopf bifurcation point.

We begin first with the period of the pulsations. Because of the general tendency of the period of delayed systems to lock to the delay time, we see that the period is relatively constant over the full range of r. Roughly speaking, our analysis is within approximately 2% of the numerically computed value for lower values of r and, as expected, gets much better as r is increased. In addition, there is excellent agreement in how the period changes for changes in β and τ. Thus, we have analytically derived the locking phenomena that to leading order is $P \approx 2\tau$ as well as reasonably captured the $O(\epsilon)$ correction.

We also have very good fit between the analytical and numerical values of the peak susceptible population x_f, as shown in Fig. 6a and d. It is clear that our result for x_f as a function of β and τ does an excellent job of matching the numerical curves as these parameters are varied. In Fig. 9 we have flattened the curves by adding the function $1-r$ so that we could expand the x_f axis and still view the full range of r; as described in Sect. 6.2, we capture the intersection of the curves for different β quite accurately.

The fit for y_f is not as accurate as that for x_f and P_f. The order of magnitude of y_f is approximately correct, but the slope as a function of r does not match as well as it does for the other quantities. However, the change in y_f as a function of β and τ is accurately captured.

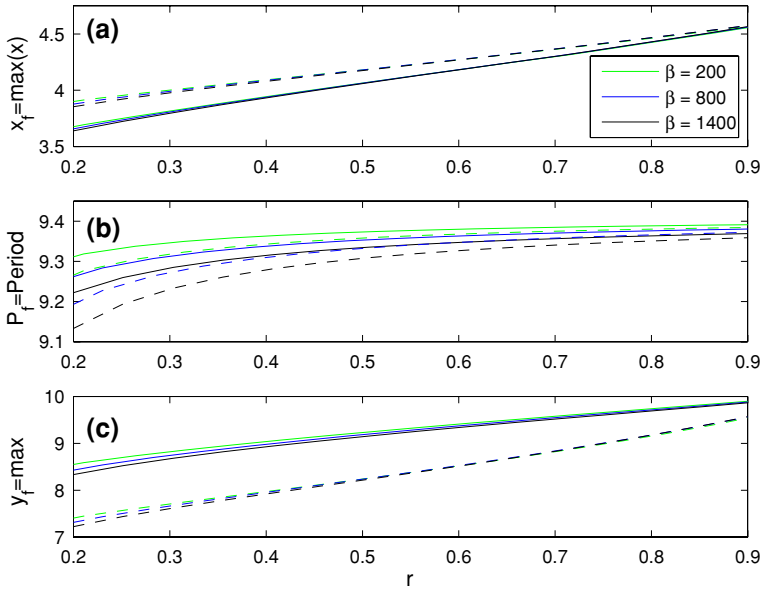

Fig. 10 **a–c** Same as in Fig. 6a–c but with $\epsilon = 0.014$

In Fig. 10 we use the same parameter values as in Fig. 6a–c but decrease ϵ to 0.01, corresponding to weaker diseases. The figure shows the quality of the leading order map results because the $O(\epsilon)$ correction terms are made smaller.

In summary, Eqs. 95 and 96 can be used to predict the magnitudes and periods of periodic epidemics as a function of the three parameters r, β and τ. Changes in the three basic quantities x_f, P_f and y_f as functions of the parameters are particularly well described.

6.6 Longer delays

In the derivation of our map we assumed that the delayed pulse that occurred at the time $t_d \equiv t_0 + \tau$ was the delayed version of the pulse at t_0 (see Fig. 7). However, for larger delays the delayed pulse that occurs at $t_d \in [t_0, t_1]$ may have originated from a pulse N oscillations in the past at t_{0-N}. If this is the case, then by definition $t_d = t_{0-N} + \tau$. Or, with reference to the pulse that started the current interval, we have that

$$t_d = t_0 + \left(\tau - \sum_{j=1}^{N} P_{0-j} \right), \tag{97}$$

where the P_{0-j} are the time intervals between the previous pulses. If the oscillations are periodic such that the time period between all pulses is equivalent, then

$$t_d = t_0 + (\tau - N P_f). \tag{98}$$

If we re-derive the map from scratch, taking into account the possibility that the delayed pulse occurred N periods in the past, and then look for periodic solutions as fixed points of this new map, we obtain new results for x_f and P_f as

$$\begin{aligned} x_f &= \frac{2\tau - d_1(1-r)}{2 + d_2(1-r)} - \frac{\epsilon\tau^2}{6r(1+2N)^3}\left[(2+r)\eta + 2(1-r)b - 4bNr\right], \\ P_f &= \frac{2\tau}{1+2N} - \frac{\epsilon\tau^2}{3r(1+2N)^3}\left[(2+r)\eta + 2b\right], \end{aligned} \tag{99}$$

with y_f again given by Eq. 96. The result is that the observed period *locks* to some fraction of the delay. Similar calculations have been made for lasers with delayed feedback by Carr (2003) and Grigorieva et al. (1992). When we compare the results of Eqs. 99 to numerical simulations using long delays (e.g. $\tau = 21\pi/2$ and $31\pi/2$) we see the same excellent qualitative fit as described in the previous section with shorter delays.

For our system, we know from the linear stability analysis that the periodic oscillations that appear at the Hopf bifurcation point have a period that is approximately 2π (see Eq. 17). In general, the integer N that is selected is such that P_f in Eq. 99 is approximately 2π. For example, we found that when $\tau = 21\pi/2$ then $N = 4$, while when $\tau = 31\pi/2$ then $N = 7$.

7 Discussion

In the first part of our discussion we will summarize the analysis and basic mathematical results presented in this paper. We will follow this with a discussion that focuses on the epidemiological interpretation of our results and their consequences.

7.1 Summary of analysis and results

It is well known that introducing delays into a system can lead to oscillatory behavior. However, the mathematical analysis of delay systems has traditionally been limited to linear stability analysis or existence proofs. Crudely speaking, the former can identify for what parameter values the endemic steady state is unstable. The latter can often provide parameter ranges or bounds such that periodic solutions must exist. In the present paper we use asymptotic methods to determine how the amplitude and period of oscillations functionally depend on the physical parameters; our primary control parameters are the delay time and the fraction of individuals who become resusceptible. The analysis methods are general in that they can be applied to other delay systems. More specifically, our results can be easily extended to the study of related systems such as SEIR models, more general competition models, or extensions of the current model, for example to the consideration of multistrain diseases (the latter will be the focus of a future manuscript). Finally, the analytical results are particularly useful in guiding numerical simulations when a quantity of interest depends in a nontrivial way on a combination of parameters. For example, having a functional relationship

between the amplitude and the parameters identifies which are the primary parameters of interest and how sensitive that dependence is, both of which can avoid the need for guesswork when starting simulations.

For all of our results, we validate numerical computations against analytical approximations and vice versa. It should be noted that the numerical simulation of DDEs is not as straightforward as that for ODEs due to having to account for the system's history (Thompson and Shampine 2006). In addition, while the basic algorithms for numerical continuation of ODEs are mature (Doedel and Oldeman 2007), continuation algorithms for DDEs (Engelborghs et al. 2001) are a much newer development and the results are not always as robust. Furthermore, the pulsations followed by long quiescent period in our problem require careful monitoring of the accuracy of results. Thus, whenever possible we consider standard operating procedure for having a complete understanding of a problem to be to use both approaches in tandem.

After we rewrite our SIRS model in non-dimensional form, we perform a linear stability analysis of the non-zero endemic state. Exact solutions of the characteristic equation are impossible to obtain. However, using perturbation methods we can determine the parameter values, including delay times, for which the endemic state is unstable with growing oscillations via a Hopf bifurcation. We find that depending upon the delay, the first (lowest r) bifurcation is to either a natural mode, whose frequency is close to the quasifrequency of the system without delay, or a delay mode, whose frequency is fixed by the delay time.

The periodic epidemics that arise when the endemic steady state is unstable to a natural mode can be described using results derived by Pieroux et al. (1994). We have specialized their results for our system and found that they well describe the amplitude of the oscillations. In addition, the change from supercritical to subcritical bifurcations near the minimum of the neutral stability curve is predicted.

To describe the periodic epidemics due to the delay modes we need to allow the resusceptible fraction r to be $O(1)$ such that the leading-order problem contained the delay. However, by looking for periodic solutions we are able to use the method of multiple scales to derive a bifurcation equation that describes the amplitude of the oscillations. We find the delay modes with frequency $\omega = m\pi/\tau$, m an even integer, bifurcate subcritically such that there is an interval of r where oscillations coexist with the endemic state. On the other hand, if m is odd the bifurcation is supercritical.

To describe the pulsating epidemics that occur when the resusceptible fraction is $O(1)$, we derive a map based on solutions constructed using matched asymptotics. More specifically, we patch together solutions that are individually determined from approximate equations in subintervals of the full period, e.g., when the epidemic is large, when the epidemic is small, or when the "delayed epidemic" is large. This technique has been previously used to describe pulsating lasers, both with delay and without (see Schwartz and Erneux 1994; Carr 2003; Grigorieva et al. 1992; Carr et al. 2000). However, in these earlier analyses the pulse width was much smaller than the overall period and could be ignored. This was not true in the present problem, and to obtain a good fit between the analytical and numerical results, we were required to find an approximate expression for the pulse width and include its effect in the leading order map. While the calculation is rather lengthy, the final equations that describe the amplitude and period of the epidemics are fairly simple.

Qualitatively, our results are as follows. As described above, when r is increased beyond the Hopf bifurcation point, the populations begin to exhibit harmonic oscillations with the period close to 2π. As the resusceptible fraction is increased, the oscillations in the infectious population become pulsating, the susceptible population exhibits a triangular shape, and both increase in amplitude and period. The triangular shape is due to the fast depletion of susceptible individuals when there is an epidemic, and then the slow resupply of susceptible individuals after the epidemic. As r is increased further, the period locks to approximately $2\tau/(1+2N)$ for some integer N, while the peaks in oscillations continues to grow.

For r away from the Hopf bifurcation, the susceptible population transitions from being a triangular shape to more of a square shape. Without delay, as the amplitude of the epidemic peaks becomes increasingly large, the susceptible population takes a long time to recover, and the period between epidemics becomes increasingly long (Schwartz and Erneux 1994). However, in the present problem the susceptible population is resupplied at time τ after the original epidemic. This causes a jump in the susceptible population leading not only to the square shape, but also a shorter duration of time until the next epidemic. Indeed, this is the physical interpretation of the locking of the period to the delay time.

For values of the delay corresponding to the results shown in Figs. 2–5, the periodic oscillations that appear at the primary bifurcation are stable as r is increased to $r_{\max}$. More specifically, we have not observed quasiperiodic oscillations or chaos for these values of the delay. However, for longer values of the delay, on the order of the lifetime of the individual ($O(1/\mu)$), it is possible to observe more complex oscillations as r is increased (Pieroux et al. 2000). We have not explored these solutions in any detail.

7.2 Epidemiological interpretation

Temporary immunity plays a role in the spread of diseases such as cholera, pertussis, influenza and malaria. The DDEs we study include a fixed delay corresponding to all individuals retaining immunity for the same amount of time. We assume that $\mathcal{R}_0 > 1$ such that if the delay is zero the endemic steady state is stable. Our analysis describes periodic epidemics that appear via a Hopf bifurcation of the endemic steady state, where we have found that the critical value of the resusceptible fraction $r = r_h$ depends upon the immunity time. For intervals of the immunity time such that the natural mode is the first to bifurcate, we have that $r_h \ll 1$. Thus, only a small fraction of the population needs to lose its immunity to generate recurrent epidemics in the whole population. On the other hand, there is a minimum immunity time necessary such that periodic epidemics will occur; that is, the delay must be greater than $\tau_{\min}$ for there to be a Hopf bifurcation.

The immunity time also determines if the bifurcation is supercritical or subcritical. In the case of the former, as the resusceptible fraction increases the size of the epidemics increases from zero monotonically. In contrast, if the bifurcation is supercritical then the system will jump from the endemic steady state to large ($O(1)$) amplitude epidemics. From the point of view of the epidemiologists in the field, this implies that it is possible for large amplitude epidemics to occur with essentially little warning. This is because, in general, no disease is modeled well enough to precisely quantify and

measure r_h. If the bifurcation is supercritical, then the small oscillations that occur for $r > r_h$ could possibly be observed before r increases such that large amplitude oscillations are generated. However, if the bifurcation is subcritical, then large oscillations could occur before it is even known that the system is close to $r = r_h$.

The subcritical bifurcation also creates an interval of bistability between the endemic steady state and the recurrent epidemics. If we assume that the system is initially in the endemic steady state, then, in general, a disturbance on the order of the size of the amplitude of the oscillatory state is needed to kick the system into the latter's basin of attraction. Outside of some catastrophic event, we would expect such large disturbances to be rare in real populations. However, if oscillations were initiated, either via a catastrophe or if $r > r_h$, then the bistability could make it equally difficult to stop the recurrent epidemics. Either another large disturbance would be needed to send the system back to the endemic state's basin of attraction, or the resusceptible fraction must be reduced to less than the lefthand limit point of the bifurcation curve. In the latter case, if the interval of bistability were small, as in Fig. 4, then the reduction in r that would be needed is small. However, for other values of the delay, such as in Fig. 5b and c, the interval of bistability is large and the resusceptible fraction would need to be reduced substantially to eliminate the epidemics.

Our results do an excellent job of predicting how the oscillations will respond to changes in the physical parameters such as β, ϵ (μ and γ) and τ. Figures 6, 10 all show that the analytical results accurately predict how much the peak values of the susceptible and infectious populations and the period will change, given a change in the parameters. For example, increasing the transmission coefficient β causes a *decrease* in the severity and period of the epidemics; this is because increasing β increases the endemic number of infectious individuals and there are fewer available susceptible individuals for "epidemic" oscillations about the endemic state.

Equations 95 indicate, and Figs. 6, 10 confirm, that changes in the delay time lead to large ($O(1)$) changes in the amplitude and period of the epidemic when the latter is pulsating. Specifically, larger delay times lead to longer periods between epidemics and larger epidemics. This is because a longer delay time allows for a larger number of individuals to populate the temporarily immune (removed) class before being re-injected into the susceptible population. The large influx of new susceptible individuals allows the system to support a large new epidemic spike of infectious individuals.

In general, the dependence of the properties of the epidemic on the delay time can be understood by examining the limiting cases of $\tau \to 0$ and $\tau \to \infty$. When $\tau \to 0$, individuals become susceptible to re-infection immediately after recovery and our SIR *with delay* model approaches an SIS model. With $\tau \to \infty$, all individuals gain permanent immunity and our SIR model with partial temporary immunity becomes a classical SIR model with permanent immunity for all removed individuals.

Finally, we note that as the peak value of the epidemic grows, the post-epidemic level of the infectious population decreases. In Fig. 7 this corresponds to $y(t)$ getting closer and closer to the invariant line $y = -1$. For low numbers of infectious individuals, stochastic effects become important such that extinction of the disease is possible, see (Dykman et al. 2008). Thus, in general, as the resusceptible fraction is increased, leading to higher peak values of the infectious population, there is an

increased probability of disease extinction between the infectious peaks. Using the results of Sect. 6 we could determine a relationship that provides the minimum value of $y(t)$ as a function of the parameters and the peak value, and then relate that to probability of extinction for different noise levels. However, the detailed analysis of that issue is beyond the scope of the present paper.

References

Anderson RM, May RM (eds) (1991) Infectious diseases of humans: dynamics and control. Oxford University Press, Oxford, UK

Brauer F, Castillo-Chavez C (2001) Mathematical models in population biology and epidemiology. Springer, New York

Bestehorn M, Grigorieva EV, Haken H, Kaschenko SA (2000) Order parameters for class-B lasers with a long time delayed feedback. Phys D 145:110–129

Carr TW (2003) Period locking due to delayed feedback in a laser with saturable absorber. Phys Rev E 68:026212

Carr TW, Billings L, Schwartz IB, Triandaf I (2000) Bi-instability and the global role of unstable resonant orbits in a driven laser. Phys D 247:59–82

Carr TW, Schwartz IB, Kim M-Y, Roy R (2006) Delayed-mutual coupling dynamics of lasers: scaling laws and resonances. SIAM J Appl Dyn Syst 5:699–725

Chow S-N, Diekmann O, Mallet-Paret J (1985) Stability, multiplicity and global continuation of symmetric periodic solutions of a nonlinear Volterra integral equation. Jpn J Appl Math 2:433–469

Cooke K, Van Den Driessche P (1996) Analysis of an SEIRS epidemic model with two delays. J Math Biol 35:240–260

Cooke K, Van Den Driessche P, Zou X (1999) Interaction of maturation delay and nonlinear birth in population and epidemic models. J Math Biol 39:332–352

Diekmann O, Montijn R (1982) Prelude to Hopf in an epidemic model: analysis of a characteristic equation associated with a nonlinear volterra integral equation. J Math Biol 14:117–127

Doedel EJ, Oldeman BE (2007) AUTO-07P: Continuation and bifurcation software for ordinary differential equations. California Institute of Technology and Concordia University

Driver RD (1977) Ordinary and delay differential equations. Springer, New York

Driver RD, Sasser DW, Slater ML (1973) The equation $x'(t) = ax(t) + b(t - \tau)$ with 'small' delay. Am Math Mon 80:990–995

Dykman MI, Schwartz IB, Landsman AS (2008) Disease extinction in the presence of random vaccination. Phys Rev Lett 101:078101

El'sgol'ts LE, Norkin SB (1973) Introduction to the theory and application of differential equations with deviating arguments (translated by J.L.Casti). Academic Press, New York

Engelborghs K, Luzyanina T, Samaey G (2001) DDE-BIFTOOL v. 2.00 user manual: a Matlab package for bifurcation analysis of delay differential equations, Technical Report TW-330, Department of Computer Science, K.E.Leuven, Leuven, Belgium

Grigorieva EV, Kahchenko SA, Loika NA, Samson AM (1992) Nonlinear dynamics in a laser with a negative delayed feedback. Phys D 59:297

Hethcote HW (2000) The mathematics of infectious diseases. SIAM Rev 42:599–653

Hethcote HW, Stech HW, Van Den Driessche P (1981) Nonlinear oscillations in epidemic models. SIAM J Appl Math 40:1–9

Kevorkian J, Cole JD (1996) Multiple scale and singular perturbation methods. Springer, New York

Kim MY, Roy R, Aron JL, Carr TW, Schwartz IB (2005) Scaling behavior of laser population dynamics with time-delayed coupling: theory and experiment. Phys Rev Lett 94:088101

Luzyanina T, Roose D, Bocharov G (2005) Numerical bifurcation analysis of immunological models with time delays. J Comput Appl Math 184:165–176

Pieroux D, Erneux T (1996) Strongly pulsating lasers with delay. Phys Rev A 53:2765–2771

Pieroux D, Erneux T, Otsuka K (1994) Minimal model of a class-B lasers with delayed feedback: Cascading branching of periodic solutions and period doubling bifurcation. Phys Rev A 50:1822–1829

Pieroux D, Erneux T, Gavrielides A, Kovanis V (2000) Hopf bifurcation subject to a large delay in a laser system. SIAM J Appl Math 61:966–982

Schwartz IB, Erneux T (1994) Subharmonic hysteresis and period doubling bifurcations for a periodically driven laser. SIAM J Appl Math 54:1083–1100
Schwartz IB, Smith HL (1983) Infinite subharmonic bifurcation in an SEIR epidemic model. J Math Biol 18:233–253
Thompson S, Shampine LF (2006) A friendly Fortran DDE solver. Appl Numer Math 56:503–516