J. Math. Biol. 45, 471–510 (2002)
Digital Object Identifier (DOI):
10.1007/s00285-002-0163-9

Mathematical Biology

J.H.P. Dawes · J.R. Gog

The onset of oscillatory dynamics in models of multiple disease strains

Received: 5 December 2001 / Revised version: 5 May 2002 /
Published online: 17 October 2002 –

Abstract. We examine a generalised SIR model for the infection dynamics of four competing disease strains. This model contains four previously-studied models as special cases. The different strains interact indirectly by the mechanism of cross-immunity; individuals in the host population may become immune to infection by a particular strain even if they have only been infected with different but closely related strains. Several different models of cross-immunity are compared in the limit where the death rate is much smaller than the rate of recovery from infection. In this limit an asymptotic analysis of the dynamics of the models is possible, and we are able to compute the location and nature of the Takens–Bogdanov bifurcation associated with the presence of oscillatory dynamics observed by previous authors.

1. Introduction

One of the major challenges in the modelling of infectious diseases is to capture the dynamics of multiple disease strains [11]. Many pathogens of importance have several different antigenic variants present in a host population simultaneously. The classic example is influenza, where there are several circulating subtypes currently prevalent in the human population, with many minor variants within each subtype [3]. Other important examples include meningitis [15], dengue [7] and malaria [14]. In this paper we consider multiple disease strains that are present in a large host population and which interact via host cross-immunity. The key idea is that infection with one strain confers on the host individual some partial cross-immunity to other strains. Cross-immunity is included in different ways in different models, but the central idea is the same: infection with one strain of the disease produces a lasting immune memory in the host which acts to protect against subsequent infections by other strains.

Previously, several authors have formulated multiple strain models with cross-immunity, most notably [1, 8, 10, 13]. All of these are extensions of essentially the

J.H.P. Dawes: DAMTP, University of Cambridge, Silver Street, Cambridge CB3 9EW, UK. e-mail: J.H.P.Dawes@damtp.cam.ac.uk

J.R. Gog: Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK. e-mail: julia@zoo.cam.ac.uk

Keywords or phrases: Infection – Pathogen – Epidemiology – Multiple strains – Cross-immunity – Oscillations – Dynamics – Bifurcations

same single strain system, given by

$$\dot{I} = \beta\theta I - \nu I - bI, \tag{1}$$

$$\dot{\theta} = b - b\theta - \beta\theta I, \tag{2}$$

where I is the proportion of the host population that is infected, θ is the proportion of susceptible individuals, b is the host birth and death rate (these are equal as we suppose the host population size to be constant and we work with host population proportions) and β is the transmission coefficient. To simplify these ODEs we re-scale time by a factor of b (we view the dynamics on the time-scale of the typical host lifetime) and we work with the so-called 'force of infection', $\Lambda \equiv \beta I/b$ rather than the proportion of infecteds. The force of infection should be thought of as the rate at which susceptible individuals become infected. After this change of variables equations (1)–(2) become

$$e\dot{\Lambda} = \Lambda(r\theta - 1),$$

$$\dot{\theta} = 1 - \theta - \theta\Lambda,$$

where the basic reproduction ratio $r \equiv \beta/(b + \nu)$ should be thought of as the expected number of people a single individual would infect in an otherwise susceptible population. The timescale parameter $e = b/(b+\nu)$ is the ratio of a typical infectious period to a typical host lifetime.

Each of [1,8,10,13] extend this description to more than one disease strain and include cross-immunity. In this way we are led to define a force of infection Λ_i for each strain, $i = 1, 2, ..., n$. The susceptible proportion of the population for each strain, θ_i, becomes more complicated; this is where the interaction between strains is concentrated. Each of the papers mentioned above introduces additional variables that are now required to keep track of what is happening to the host population. There are at least two ways to introduce these new variables. 'History-based' models describe host individuals by labelling them with the collection of those strains they have previously been infected by and are therefore immune to. The host may have some immunity to other strains if it has similar strains in its history of infection. In contrast, 'status-based' models describe the host individual according to their current immune status. In its simplest form this is a list of the strains to which the host is currently immune, regardless of how the host arrived at this immune memory. We will be using the same notation for both models; S_J where J is a subset of the collection of available strains, though its interpretation subtly differs in the two cases. For history-based models, J is the set of previous infections experienced by a proportion S_J of the host population. For status-based models, J is the set of strains a proportion S_J of hosts currently has immunity to.

A further difference between models is in the way cross-immunity works. Cross-immunity could affect the susceptibility of hosts (the rate at which they are infected) or the transmissibility of future infections (the rate at which hosts infect others). These may result in different dynamics; in the latter case host individuals may still be gaining further immunity to other strains.

There are of course many similarities between these different models in terms of how they behave. However there are differences which have not been adequately

explained, nor ascribed to differences in the assumptions of a particular model in a satisfactory way. These differences are most apparent in comparisons of systems of four strains, the subject of this paper. Gupta *et al.* [13] considered the set of strains to be generated from all possible combinations of alleles at several polymorphic antigenic sites of the pathogen. If there are two alleles possible at each of two difference loci, then there are four possible strains. Also, Gomes *et al.* [10] indicated that the dynamics of four strains appears to be sufficiently complex to be of great value in describing and understanding the dynamics of a larger number of strains.

As we might expect, all existing models of multiple strains behave the same at extreme values of cross-immunity. When infection with one strain gives a host very little immunity to infection by other strains, the model reaches an equilibrium where all strains invade the host population and reach equal levels. Contrastingly, when infection with one strain gives a host a high level of immunity to subsequent infection by other strains, an equilibrium is reached where only a subset of the available strains have invaded the population. The dynamics at intermediate levels of cross-immunity is more complex. Numerical solutions of various four-strain models (those of Gomes *et al.* [10], Gupta *et al.* [13] and Andreasen *et al.* [1]) indicated that long-term stable solutions for intermediate values of a cross-immunity parameter are oscillatory. In contrast, there are apparently no oscillations in the status-based model described by Gog & Swinton [8] for any parameter values. It is this key difference between the behaviour of existing models which has motivated the work described in this paper.

In a sense this paper continues the discussion of points raised by Gog & Swinton [8]; understanding the dynamics of multiple strains has a clear biological relevance, and the models of these interactions are both mathematically and biologically complex. In general we hope first to gain a clear mathematical understanding of the models, and then to translate this into biological insights into how different sets of initial assumptions influence the resulting model dynamics.

The contents of the paper are as follows. In section 2.1 we describe the construction of these four-strain models in more detail, and define the different assumptions behind each of the models discussed later. In section 2.2 we introduce the generalised model for the dynamics of four strains, and discuss its steady and oscillatory dynamics near $r = 1$, using an asymptotic analysis in the limit of small e to make analytic progress. The details of the calculation are relegated to the Appendix. Section 3 discusses four specific models that have been discussed by other authors. Using the analysis of the general model we are able to explain their results in a complete and consistent fashion, and to compare the dynamics of different models. This analysis provides a level of mathematical detail which we feel has previously been lacking. We conclude in section 4.

2. The generalised model

2.1. Preliminaries

The most basic common feature of the models we consider is that they have four different disease strains, thought of as arranged in a circle, see figure 1. There is

symmetry in the model construction; each strain interacts with the strains adjacent to and opposite it, in the same way and all strains are equally virulent (they share the same transmission coefficients and recovery rates). Then the model is symmetric under the action of the group D_4 (the symmetries of a square), generated by the permutation operations

$$\rho : \{1, 2, 3, 4\} \to \{2, 3, 4, 1\}, \tag{3}$$

$$m_x : \{1, 2, 3, 4\} \to \{3, 2, 1, 4\}, \tag{4}$$

acting on the set of strains $\{1, 2, 3, 4\}$. Infection with any strain confers total cross-immunity to subsequent infection with that strain. Hence, no individual can be infected with the same strain twice. In addition, each strain confers some partial protection to adjacent strains. There is little or no cross-immunity between strains that are opposite each other on the circle; hence there is a natural pairing of strains in this system. Strains 1 and 3 are each in competition with strains 2 and 4, and so are indirectly mutually beneficial. Hence it is natural to think of strains 1 and 3 as forming a pair, and 2 and 4 another. We now argue that strains within one pair will tend to come into an equilibrium where the two strains are equally prevalent in the population.

Even without directly considering any mathematical system, it is easy to imagine what will happen at high and low values of cross-immunity. At one extreme, when there is little interaction between strains, all strains would tend to the same levels of prevalence so the long-term stable equilibrium state would be totally symmetric. That is, assuming the parameters were such that each strain could persist in the absence of the others, we would expect them all to be present at equal levels in the population. At the other extreme, if there were high levels of competition between strains, then the persistent presence of one strain would enable it's opposite number to be able to invade. This can be argued as follows. Imagine strains 1 and 3 are present in the population, and are at equilibrium. Then they both generate a cross-immunity to strain 2 in the population. At high cross-immunity, each of strains 1 and 3 are producing almost as much cross-immunity to 2 as they are to themselves. Faced with this double inhibition by adjacent strains, strain 2 cannot

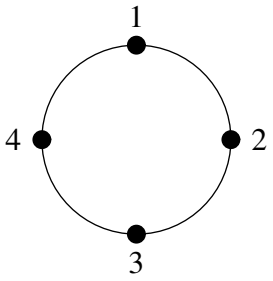

Fig. 1. The four strains interact as if equally spaced around a circle; each strain interacts more with its adjacent neighbours than with the strain opposite.

invade. The same holds for strain 4. So the high level of competition keeps the other pair excluded from the population. Exactly which pair is present and which is absent will depend on initial conditions; since these two equilibrium states are related by the symmetry of the system they are equivalent dynamically. In later discussions we will refer to such an equilibrium state, with two strains persisting and two excluded, as an 'edge' equilibrium, for geometrical reasons which should become clear.

All the models contain two basic (positive) parameters: the basic reproduction ratio r, and the ratio of timescales e defined in the introduction. Each model also contains parameters defining the cross-immunity structure. We note that two models (those discussed by [1] and [8]) use only one parameter, σ, to define cross-immunities. For consistency in discussing these particular models, we interpret the cross-immunity parameter σ as defined in [1], which is the reverse of the definition in [8]: $\sigma = 1$ corresponds to no cross-immunity between adjacent strains, and $\sigma = 0$ corresponds to complete cross-immunity. The cross-immunity structure of the model is also affected by assumptions concerning the accumulation of cross-immunity in individuals. One of two formulations is used in each of the specific models we analyse. As the name suggests, in the 'product' formulation the cross-immunity a host develops to a given strain, as a result of repeated infections, increases with each infection, by a factor depending on how closely related each strain is to the given strain. In the 'minimum' formulation the host's susceptibility is taken to be the least that would result from each previous infection, taken one at a time.

Finally, the models that we consider do not include 'removal of infecteds'. We do not assume that an individual cannot be simultaneously infected by more than one strain, and we do not omit them from the susceptible population while they are infected with another strain. In other words, multiple simultaneous infections are possible. In the the case of small e, corresponding to relatively short-lived infections, whether or not we make this assumption makes little difference. However it greatly simplifies the construction of the generalised model. This assumption is discussed further in section 3.2, in relation to the model analysed by Andreasen et al [1].

2.2. *Model definition*

In this section we present a very general model for the dynamics of four strains. As discussed in the introduction, the susceptible host population is divided into 2^n classes S_J where the labelling set J is always a subset of $\{1, 2, 3, 4\}$, omitting the brackets for ease of notation. Movement between these classes is at rates given by the non-negative coefficients a_j, b_j, c_j, d_j, e_j and f_j that are combinations of other parameters which define the cross-immunity structure. To define the model it is enough to list the ODEs that give the dynamics of Λ_1 and the classes S_J for subsets J which are not related by symmetries. Equations for the remaining variables $\Lambda_2, \ldots, \Lambda_4$ and the other S_J can be obtained by applying the permutation symmetries (3) and (4).

The equations for the generalised multiple strain model are then

$$e\dot{\Lambda}_1 = \Lambda_1(r\theta_1 - 1), \tag{5}$$

where we have defined

$$\theta_1 = S_\emptyset + \tau_1(S_2 + S_4) + \tau_2 S_3 + \tau_3(S_{23} + S_{34}) + \tau_4 S_{24} + \tau_5 S_{234}, \tag{6}$$

and

$$\begin{aligned}
\dot{S}_\emptyset &= 1 - S_\emptyset - (\Lambda_1 + \Lambda_2 + \Lambda_3 + \Lambda_4)S_\emptyset, && (7)\\
\dot{S}_1 &= -S_1 + a_1 S_\emptyset \Lambda_1 - b_0(\Lambda_2 + \Lambda_4)S_1 - c_0\Lambda_3 S_1, && (8)\\
\dot{S}_{12} &= -S_{12} + a_2 S_\emptyset(\Lambda_1 + \Lambda_2) + b_1(\Lambda_1 S_2 + \Lambda_2 S_1) - d_0(\Lambda_3 + \Lambda_4)S_{12}, && (9)\\
\dot{S}_{13} &= -S_{13} + c_1(\Lambda_1 S_3 + \Lambda_3 S_1) + a_3 S_\emptyset(\Lambda_1 + \Lambda_3) - e_0(\Lambda_2 + \Lambda_4)S_{13}, && (10)\\
\dot{S}_{123} &= -S_{123} + a_4 S_\emptyset(\Lambda_1 + \Lambda_3) + a_5 S_\emptyset \Lambda_2 + b_2(S_1 + S_3)\Lambda_2\\
&\quad + c_2(S_1\Lambda_3 + S_3\Lambda_1) + b_3 S_2(\Lambda_1 + \Lambda_3) + d_1(\Lambda_1 S_{23} + \Lambda_3 S_{12})\\
&\quad + e_1\Lambda_2 S_{13} - f_0\Lambda_4 S_{123}, && (11)\\
\dot{S}_{1234} &= -S_{1234} + a_6 S_\emptyset(\Lambda_1 + \Lambda_2 + \Lambda_3 + \Lambda_4)\\
&\quad + b_4[(S_1 + S_3)(\Lambda_2 + \Lambda_4) + (S_2 + S_4)(\Lambda_1 + \Lambda_3)]\\
&\quad + c_3(S_1\Lambda_3 + S_3\Lambda_1 + S_2\Lambda_4 + S_4\Lambda_2)\\
&\quad + d_2[(\Lambda_3 + \Lambda_4)S_{12} + (\Lambda_4 + \Lambda_1)S_{23} + (\Lambda_1 + \Lambda_2)S_{34} + (\Lambda_2 + \Lambda_3)S_{14}]\\
&\quad + e_2[(\Lambda_2 + \Lambda_4)S_{13} + (\Lambda_1 + \Lambda_3)S_{24}]\\
&\quad + f_0(\Lambda_1 S_{234} + \Lambda_2 S_{134} + \Lambda_3 S_{124} + \Lambda_4 S_{123}). && (12)
\end{aligned}$$

The negative terms represent movement out of a class as a host becomes infected with another strain, and the positive terms are influxes of these hosts into the new class. Note that equation (12) is not needed; as S_{1234} does not appear in any other equation, (12) decouples and will be ignored in what follows. The positive coefficients τ_j are usually thought of as taking values less than one, and express the idea of reduced transmissibility. The other coefficients independently allow cross-immunity to reduce susceptibility. The coefficients are constrained by the following relations

$$\begin{aligned}
a_1 + 2a_2 + a_3 + 2a_4 + a_5 + a_6 &= 1, && (13)\\
b_1 + b_2 + b_3 + b_4 &= b_0, && (14)\\
c_1 + 2c_2 + c_3 &= c_0, && (15)\\
d_1 + d_2 &= d_0, && (16)\\
e_1 + e_2 &= e_0, && (17)
\end{aligned}$$

which ensure that the flows between host classes balance and the total host population size remains constant.

2.3. *Steady dynamics*

When $r < 1$, each infected individual gives rise to, on average, less than one new case of infection. After long times, therefore, we expect that the disease will not be able to persist in the host population, and so will die out. This behaviour corresponds to the 'trivial' solution with $\Lambda_j = 0$, $S_\emptyset = 1$ and $S_J = 0$ for all other J being an attracting equilibrium.

For $r - 1 > 0$ and small, we expect the strains will only be able to invade a small proportion of the host population; steady states will have Λ_j and S_J small, except for $S_\emptyset$ which will remain close to 1. In bifurcation-theoretic terms, as r increases through 1 four real eigenvalues of the Jacobian (linearisation) matrix around the trivial solution pass through zero. Near this bifurcation we can find a smaller set of ODEs which will capture all the dynamics near $r = 1$; this is a centre manifold reduction. ODEs for the four variables Λ_j describe the dynamics on the centre manifold, and the remaining variables S_J can be systematically eliminated. Formally we use $r - 1$ as a small parameter and note that this is of the same asymptotic size as the Λ_j and S_J, except for $S_\emptyset$ for which $r - 1 \sim 1 - S_\emptyset$. We define the new variable $U = 1 - S_\emptyset$. The idea is that from inspection of the full ODEs we observe that U and all the other S_J are decaying exponentially fast towards zero, whereas this is not the case for the Λ_j since $r - 1$ is small. After this exponential decay, $\dot{S}_J \approx 0$ and $\dot{U} \approx 0$, and so we can re-arrange these equations to solve for the S_J in terms of the Λ_j and substitute these leading-order approximations into the expressions for θ_j and hence the equations for the dynamics of the Λ_j. This approach is often referred to as 'adiabatic elimination' and can be carried out in a formal asymptotic manner. It is exactly equivalent to a centre manifold reduction performed in the the usual manner. We find $S_\emptyset = 1 - \Lambda_1 - \Lambda_2 - \Lambda_3 - \Lambda_4 + \mathcal{O}(2)$, $S_1 = a_1\Lambda_1 + \mathcal{O}(2)$, $S_{12} = a_2(\Lambda_1 + \Lambda_2) + \mathcal{O}(2)$, $S_{13} = a_3(\Lambda_1 + \Lambda_3) + \mathcal{O}(2)$, $S_{123} = a_4(\Lambda_1 + \Lambda_3) + a_5\Lambda_2 + \mathcal{O}(2)$ where $\mathcal{O}(2)$ denotes terms of order Λ_j^2 and higher. Substituting these (and their permutations) into the θ_j we obtain

$$e\dot{\Lambda}_1 = \Lambda_1[r - 1 - \Lambda_1 - \tilde{q}\Lambda_3 - \tilde{p}(\Lambda_2 + \Lambda_4)] + \mathcal{O}(3), \tag{18}$$

$$e\dot{\Lambda}_2 = \Lambda_2[r - 1 - \Lambda_2 - \tilde{q}\Lambda_4 - \tilde{p}(\Lambda_1 + \Lambda_3)] + \mathcal{O}(3), \tag{19}$$

$$e\dot{\Lambda}_3 = \Lambda_3[r - 1 - \Lambda_3 - \tilde{q}\Lambda_1 - \tilde{p}(\Lambda_2 + \Lambda_4)] + \mathcal{O}(3), \tag{20}$$

$$e\dot{\Lambda}_4 = \Lambda_4[r - 1 - \Lambda_4 - \tilde{q}\Lambda_2 - \tilde{p}(\Lambda_1 + \Lambda_3)] + \mathcal{O}(3), \tag{21}$$

where $\tilde{p} = 1 - \tau_1 a_1 - \tau_3 a_2 - \tau_4 a_3 - \tau_5 a_4$ and $\tilde{q} = 1 - \tau_2 a_1 - 2\tau_3 a_2 - \tau_5 a_5$. $\tilde{p}$ and $\tilde{q}$ may be interpreted within the model as follows. Imagine that we infect the whole population with one strain; $\tilde{p}$ then gives the proportion of the population who are now incapable of infecting others with an adjacent strain in the future. Similarly, $\tilde{q}$ gives the proportion of the population who are now incapable of infecting others with the opposite strain.

The system can be further simplified using the observation motivated by computing

$$e\frac{d}{dt}\log\left(\frac{\Lambda_1}{\Lambda_3}\right) = e\left(\frac{\dot{\Lambda}_1}{\Lambda_1} - \frac{\dot{\Lambda}_3}{\Lambda_3}\right) = (1 - \tilde{q})(\Lambda_3 - \Lambda_1) + \mathcal{O}(2).$$

Since $1 - \tilde{q} = \tau_2 a_1 + 2\tau_3 a_2 + \tau_5 a_5 > 0$, when $\Lambda_1 > \Lambda_3$ the RHS is negative, so Λ_1/Λ_3 decreases, and vice versa. Intuitively, infection with any strain gives more immunity to itself than to its opposite number, and affects the two strains adjacent to it equally, so we would expect the force of infection of opposite strains to settle to the same equilibrium value. We conclude that the 'diagonal' subspace where $\Lambda_1 = \Lambda_3$ and (by a similar calculation) $\Lambda_2 = \Lambda_4$ attracts all trajectories, at least close to $r = 1$. Numerical investigations of all specific models that we will discuss later show that, over a very large range of r at least, trajectories are attracted to this subspace. Using this information, we can simplify the ODEs further by looking at the dynamics within this subspace, and we employ this simplification throughout the remainder of this paper, setting $\Lambda_3 = \Lambda_1$ and $\Lambda_4 = \Lambda_2$. Within the diagonal subspace, (18) - (21) become

$$e\dot{\Lambda}_1 = \Lambda_1[r - 1 + p_1\Lambda_1 + p_2\Lambda_2] + \mathcal{O}(3), \tag{22}$$

$$e\dot{\Lambda}_2 = \Lambda_2[r - 1 + p_1\Lambda_2 + p_2\Lambda_1] + \mathcal{O}(3), \tag{23}$$

where $p_1 = -1 - \tilde{q}$ and $p_2 = -2\tilde{p}$. The coefficient $-p_1$ is a measure of the immunity given by a strain to itself and its opposite number, and $-p_2$ gives a measure of the immunity given by a strain to the pair of adjacent strains. The equations (22)–(23) have three types of equilibrium; the trivial solution $\Lambda_1 = \Lambda_2 = 0$, the two 'edge' equilibria $\Lambda_1 \neq 0$, $\Lambda_2 = 0$ and $\Lambda_1 = 0$, $\Lambda_2 \neq 0$ which are related by the symmetry ρ which has the effect of interchanging Λ_1 and Λ_2, and the 'fully-symmetric' equilibrium $\Lambda_1 = \Lambda_2 \neq 0$. By writing $x_1^2 = \Lambda_1$ and $x_2^2 = \Lambda_2$ and rescaling time by a factor of $1/2e$, we transform (22) - (23) into

$$\dot{x}_1 = x_1[r - 1 + p_1 x_1^2 + p_2 x_2^2] + \mathcal{O}(5), \tag{24}$$

$$\dot{x}_2 = x_2[r - 1 + p_1 x_2^2 + p_2 x_1^2] + \mathcal{O}(5), \tag{25}$$

where the higher-order terms are now of order x_j^5.

The transformation $x_j^2 = \Lambda_j$ should be thought of as converting the existence of invariant subspaces for (22) - (23), i.e. the fact that if $\Lambda_j = 0$ then $\dot{\Lambda}_j = 0$, into extra symmetries. The invariant subspaces exist due to the biologically-inspired restriction is that if there is none of a particular strain initially present, then none is created in the subsequent dynamics. But invariant subspaces in general are not generic features of systems of ODEs. However, the study of systems of ODEs with symmetries, but that are otherwise generic, is well advanced, see for example [9]. From this viewpoint, this change of variables is very natural. The new symmetry property here is the invariance of (24) - (25) under the transformation $(x_1, x_2) \to (-x_1, x_2)$, and similarly for x_2.

The dynamics of (24) - (25) are well-known in the bifurcation literature; they describe a steady-state bifurcation with D_4 (square) symmetry, first analysed by Swift [19]. It turns out that this bifurcation problem is closely related to the analysis of a Hopf bifurcation with $O(2)$ symmetry, discussed by, among others, Knobloch [17]. From the bifurcation theory we can assert the existence of two distinct branches of solutions (unrelated by symmetry). These are the 'edge' equilibria and the 'fully-symmetric' equilibrium referred to above. They exist in the region $r > 1$

since $p_1 < 0$ and $p_1 + p_2 < 0$. A linearised stability analysis shows that when $p_1 > p_2$ the edge equilibrium is stable and the fully-symmetric one is unstable. When $p_1 < p_2$ these stabilities are reversed. When p_1 is close to p_2 this transition is mediated by the appearance of an 'asymmetric' equilibrium where $x_1 \neq x_2$ and neither are zero. The existence of this asymmetric equilibrium is a result of the addition to (24) - (25) of the fifth and seventh-order terms needed to obtain a complete unfolding of the degenerate situation that exists when $p_1 = p_2$. The asymmetric equilibrium itself can be stable or unstable depending on the signs of these higher-order terms. The existence of this asymmetric equilibrium was reported (for different models) by Ferguson & Andreasen [6] and by Gog & Swinton [8]; in the first case it was found to be stable, and in the second case it was found to be unstable. In the unstable case this resulted in a small region of parameter space where both the edge and fully-symmetric equilibria were stable. The first case (where the asymmetric equilibrium is stable) is illustrated in figure 2(a); the second case, where the asymmetric equilibrium is unstable, is shown in figure 2(b).

In summary, the equations (24) - (25), with the addition of higher-order terms near degenerate points where $p_1 = p_2$, completely describe the dynamics of these four-strain models near the initial loss of stability of the trivial (no invasion) solution at $r = 1$. None of the steady-state dynamics depend on the value of the coefficient e; after setting all the time derivative terms to zero, e does not appear in the generalised model equations (5) - (11). In (18) - (21) the factors of e on the LHS can play no dynamic role since they can be removed by rescaling time. Moreover, the dynamics of (18) - (21) can be seen to contain no oscillations since they are a gradient flow for the potential $F(x_1, x_2, x_3, x_4)$:

$$\frac{1}{e}F(x_1, x_2, x_3, x_4) = \frac{r-1}{2}(x_1^2 + x_2^2 + x_3^2 + x_4^2) - \frac{1}{4}(x_1^4 + x_2^4 + x_3^4 + x_4^4) - \frac{q}{2}(x_1^2 x_3^2 + x_2^2 x_4^2) - \frac{p}{2}(x_1^2 + x_3^2)(x_2^2 + x_4^2)$$

where $x_j^2 = \Lambda_j$ as before. Having defined F, we find that

$$\frac{dF}{dt} = \dot{\mathbf{x}} \cdot \nabla F = |\nabla F|^2 \geq 0 \tag{26}$$

where $\mathbf{x}$ represents the vector of the x_j. Hence F increases along trajectories until an equilibrium point is reached.

2.4. Oscillatory dynamics

We now make further simplifications to the model ODEs. These scalings and changes of variables make it possible to investigate the occurrence of oscillatory dynamics in the model analytically. The complicated nature of the ODEs forces us to consider the behaviour only near $r = 1$. Due to the gradient nature of (18) - (21) (which rules out the existence of oscillations) it is clear that we must include at least some of the equations for the S_J variables in the analysis. The starting point for the analysis of

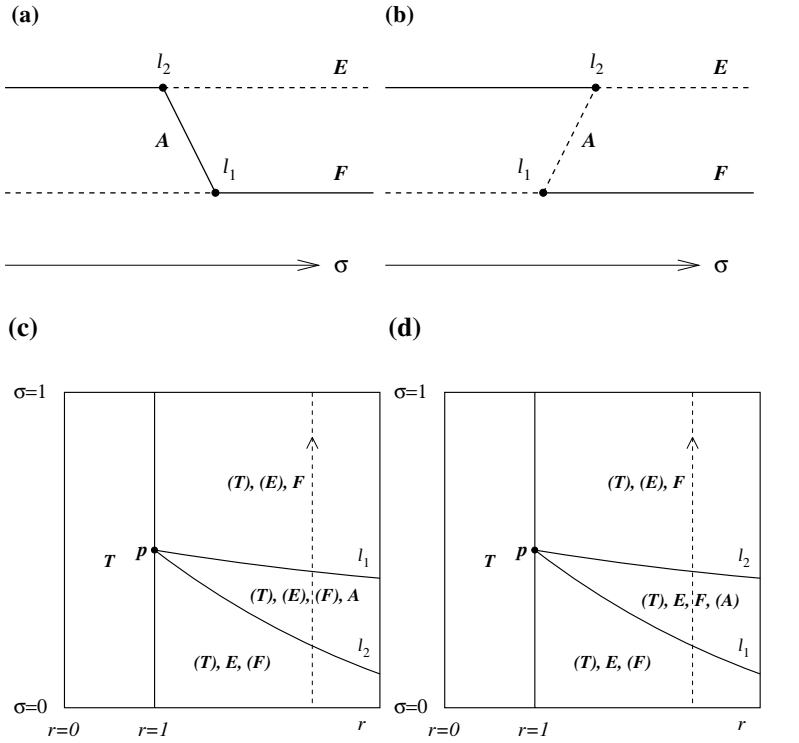

Fig. 2. Schematic illustrations of the steady dynamics near $r = 1$, assuming that $p_1 < 0$, $p_1 + p_2 < 0$ and the higher-order contributions to (24) - (25) are non-degenerate. (a) and (b), respectively, indicate the stable and unstable equilibria as the dashed vertical path in, respectively, (c) and (d), is traversed. In (c) and (d), curves starting at the point p where $p_1 = p_2$ give the boundaries of the regions of stability of the fully-symmetric equilibrium F, the edge equilibrium E and the asymmetric equilibrium A (which exists in the central region only). Brackets and dashed lines indicate unstable equilibria. In (a) and (c) the asymmetric equilibrium A is stable; in (b) and (d) it is unstable.

the remainder of this section is the full equations (5) - (11) restricted to the diagonal subspace:

$$e\dot{\Lambda}_1 = \Lambda_1(r\theta_1 - 1), \tag{27}$$

$$e\dot{\Lambda}_2 = \Lambda_2(r\theta_2 - 1), \tag{28}$$

$$\theta_1 = S_\emptyset + \tau_2 S_1 + 2\tau_1 S_2 + 2\tau_3 S_{12} + \tau_4 S_{24} + \tau_5 S_{234}, \tag{29}$$

$$\theta_2 = S_\emptyset + \tau_2 S_2 + 2\tau_1 S_1 + 2\tau_3 S_{12} + \tau_4 S_{13} + \tau_5 S_{123}, \tag{30}$$

$$\dot{S}_\emptyset = 1 - S_\emptyset - 2S_\emptyset(\Lambda_1 + \Lambda_2), \tag{31}$$

$$\dot{S}_1 = -S_1 + a_1 S_\emptyset \Lambda_1 - 2b_0 S_1 \Lambda_2 - c_0 S_1 \Lambda_1, \tag{32}$$

$$\dot{S}_2 = -S_2 + a_1 S_\emptyset \Lambda_2 - 2b_0 S_2 \Lambda_1 - c_0 S_2 \Lambda_2, \tag{33}$$

$$\dot{S}_{12} = -S_{12} + a_2 S_\emptyset(\Lambda_1 + \Lambda_2) + b_1(S_1\Lambda_2 + S_2\Lambda_1) - d_0 S_{12}(\Lambda_1 + \Lambda_2), \tag{34}$$

$$\dot{S}_{13} = -S_{13} + 2c_1 S_1 \Lambda_1 + 2a_3 S_\emptyset \Lambda_1 - 2e_0 S_{13} \Lambda_2, \tag{35}$$

$$\dot{S}_{24} = -S_{24} + 2c_1 S_2 \Lambda_2 + 2a_3 S_\emptyset \Lambda_2 - 2e_0 S_{24} \Lambda_1, \tag{36}$$

$$\begin{aligned}\dot{S}_{123} = &-S_{123} + 2a_4 S_\emptyset \Lambda_1 + a_5 S_\emptyset \Lambda_2 + 2b_2 S_1 \Lambda_2 + 2c_2 S_1 \Lambda_1 \\ &+2b_3 S_2 \Lambda_1 + 2d_1 S_{12} \Lambda_1 + e_1 S_{13} \Lambda_2 - f_0 S_{123} \Lambda_2,\end{aligned} \tag{37}$$

$$\begin{aligned}\dot{S}_{234} = &-S_{234} + 2a_4 S_\emptyset \Lambda_2 + a_5 S_\emptyset \Lambda_1 + 2b_2 S_2 \Lambda_1 + 2c_2 S_2 \Lambda_2 \\ &+2b_3 S_1 \Lambda_2 + 2d_1 S_{12} \Lambda_2 + e_1 S_{24} \Lambda_1 - f_0 S_{234} \Lambda_1.\end{aligned} \tag{38}$$

Again we have omitted the equation for S_{1234} since it decouples from the rest. Several other variables, for example S_{134} and S_{123}, obey the same equation within the 'diagonal' subspace, and so an equation for one of these (here, S_{134} in this case) can also be omitted.

First we differentiate (29) and (30) to derive evolution equations for the θ_j variables. By using the θ_j as dynamical variables we can reduce the dimension of the ODEs by two; we can use θ_1 and θ_2 to replace S_{24}, S_{234}, S_{13} and S_{123} if we assume that

$$\tau_5 e_1 + \tau_4 f_0 = 2\tau_4 e_0. \tag{39}$$

This assumption can be interpreted as assuming that the interaction between opposite strains is weak or non-existent. More precisely, (39) is the mathematical result of the following three epidemiological assumptions:

- the transmissibility of a strain is unaffected by the presence or absence of immunity to the opposite strain,
- infection by one strain cannot give the individual immunity to the opposite strain,
- immunity to one strain does not alter susceptibility to the opposite strain,

These three assumptions imply the three conditions $\tau_4 = \tau_5$, $e_0 = e_1$ and $e_0 = f_0$ respectively. Each assumption has implications for other coefficients, but these three are sufficient to satisfy (39). Assumption (39) holds for all the specific models we consider below, except that of Gomes et al. [10] (considered in section 3.4) and the general reduced transmissibility model that we consider in section 3.1, where $\tau_4 \neq \tau_5$. However, in the special case of the reduced transmissibility model studied by Ferguson & Andreasen [6], assumption (39) does hold exactly.

For convenience we now define $\phi_1 = 1 - \theta_1$, $\phi_2 = 1 - \theta_2$ and $U = 1 - S_\emptyset$. This shifts the disease-free equilibrium to the origin, for the resulting 8-dimensional system of ODEs in the variables $\{\Lambda_1, \Lambda_2, \phi_1, \phi_2, U, S_1, S_2, S_{12}\}$. The final, and crucial, change of variables is to use sums and differences of the pairs $\{\Lambda_1, \Lambda_2\}$, $\{\phi_1, \phi_2\}$ and $\{S_1, S_2\}$; we note that a similar transformation to this was used by Castillo–Chavez *et al.* [2]. We define

$$\begin{aligned}\Lambda_S &= \Lambda_1 + \Lambda_2, & \phi_S &= \phi_1 + \phi_2, & S_S &= S_1 + S_2, \\ \Lambda_D &= \Lambda_1 - \Lambda_2, & \phi_D &= \phi_1 - \phi_2, & S_D &= S_1 - S_2.\end{aligned}$$

In these co-ordinates the model ODEs become

$$e\dot{\Lambda}_S = (r-1)\Lambda_S - \frac{r}{2}(\Lambda_S \phi_S + \Lambda_D \phi_D), \tag{40}$$

$$e\dot{\Lambda}_D = (r-1)\Lambda_D - \frac{r}{2}(\Lambda_S\phi_D + \Lambda_D\phi_S), \tag{41}$$

$$\dot{\phi}_S = -\phi_S - \Lambda_S(p_1+p_2) + U\Lambda_S(p_1+f_0+p_2) - \frac{f_0}{2}(\Lambda_S\phi_S + \Lambda_D\phi_D) - p_3\Lambda_S S_S - p_4\Lambda_D S_D - 2p_7 S_{12}\Lambda_S, \tag{42}$$

$$\dot{\phi}_D = -\phi_D - \Lambda_D(p_1-p_2) + U\Lambda_D(p_1+f_0-p_2) - \frac{f_0}{2}(\Lambda_D\phi_S + \Lambda_S\phi_D) - p_5\Lambda_S S_D - p_6\Lambda_D S_S - 2S_{12}\Lambda_D(\tau_3 f_0 - \tau_5 d_1), \tag{43}$$

$$\dot{U} = -U + 2\Lambda_S - 2U\Lambda_S, \tag{44}$$

$$\dot{S}_S = -S_S + a_1\Lambda_S - a_1U\Lambda_S - b_0(\Lambda_S S_S - \Lambda_D S_D) - \frac{c_0}{2}(\Lambda_S S_S + \Lambda_D S_D), \tag{45}$$

$$\dot{S}_D = -S_D + a_1\Lambda_D - a_1U\Lambda_D - b_0(\Lambda_S S_D - \Lambda_D S_S) - \frac{c_0}{2}(\Lambda_S S_D + \Lambda_D S_S), \tag{46}$$

$$\dot{S}_{12} = -S_{12} + a_2\Lambda_S - a_2U\Lambda_S + \frac{b_1}{2}(\Lambda_S S_S - \Lambda_D S_D) - d_0 S_{12}\Lambda_S, \tag{47}$$

where the coefficients p_j correspond to the following combinations of original model coefficients:

$$\begin{aligned}
p_1 &= \tau_2 a_1 + 2\tau_3 a_2 + \tau_5 a_5 - 2,\\
p_2 &= 2(\tau_1 a_1 + \tau_3 a_2 + \tau_4 a_3 + \tau_5 a_4 - 1),\\
p_3 &= \tau_5(b_3+b_2+c_2)+\tau_4 c_1+2\tau_3 b_1+\tau_2(f_0/2-c_0/2-b_0)+\tau_1(f_0-2b_0-c_0),\\
p_4 &= \tau_5(c_2-b_3-b_2)+\tau_4 c_1-2\tau_3 b_1+\tau_2(f_0/2-c_0/2+b_0)+\tau_1(2b_0-f_0-c_0),\\
p_5 &= \tau_5(b_3-b_2-c_2)-\tau_4 c_1+\tau_2(f_0/2-c_0/2-b_0)+\tau_1(2b_0-f_0+c_0),\\
p_6 &= \tau_5(b_2-b_3-c_2)-\tau_4 c_1+\tau_2(f_0/2-c_0/2+b_0)+\tau_1(f_0-2b_0+c_0),\\
p_7 &= \tau_5 d_1 + \tau_3(f_0-2d_0).
\end{aligned}$$

From (13) it is easy to see that both p_1 and p_2 are negative when all the τ_j are less than one, which we assume is the case for this paper.

In numerical simulations, the oscillatory dynamics are seen near the fully-symmetric equilibrium, which has $\Lambda_D = \phi_D = S_D = 0$ and $\phi_S = 2(r-1)/r$ in these new co-ordinates. The simplification which these new co-ordinates produce concerns the Jacobian matrix evaluated at the fully-symmetric equilibrium; this 8×8 matrix is found to contain a 3×3 submatrix which decouples from the remaining entries. The 3×3 submatrix consists of the rows and columns corresponding to the Λ_D, ϕ_D and S_D variables. We now investigate the stability of the fully-symmetric equilibrium by computing the eigenvalues of this submatrix. We do this calculation twice; first in a non-rigorous fashion 'near $r=1$' in order to motivate the precise scalings which are applied the second time, in section 2.4.2.

2.4.1. First (unscaled) attempt at locating the Takens–Bogdanov point

Near $r=1$ we write $r = 1+\mu$ and hence $\Lambda_S = -2\mu/(p_1+p_2)$, using (22)–(23), to leading order. The 3×3 submatrix of the Jacobian matrix (evaluated at the

fully-symmetric equilibrium) becomes

$$\begin{pmatrix} 0 & \frac{\mu}{e(p_1+p_2)} & 0 \\ p_2-p_1 & -1 & \frac{2p_5\mu}{p_1+p_2} \\ a_1 & 0 & -1 \end{pmatrix}$$

which has eigenvalues given by roots of the polynomial

$$\mathcal{P}(\lambda) = \lambda^3 + 2\lambda^2 + \lambda\left[1 - \frac{(p_2-p_1)\mu}{e(p_1+p_2)}\right] + \frac{\mu}{e(p_1+p_2)}\left[p_1 - p_2 - \frac{2a_1p_5\mu}{p_1+p_2}\right].$$

A steady-state bifurcation from the fully-symmetric equilibrium occurs when $\mu = (p_1+p_2)(p_1-p_2)/2a_1p_5$ since $\mathcal{P}(\lambda)$ then has a root at zero, and so the submatrix has a zero eigenvalue there. $\mathcal{P}(\lambda)$ has a pair of purely imaginary roots, and so there is a Hopf bifurcation from the fully-symmetric equilibrium, near $r = 1$, when

$$\mu^2 - \frac{\mu(p_2-p_1)(p_1+p_2)}{2a_1p_5} + \frac{e(p_1+p_2)^2}{a_1p_5} = 0, \tag{48}$$

as long as

$$\mu(p_2-p_1) > e(p_1+p_2). \tag{49}$$

A careful analysis of the four cases corresponding to the four sign combinations of $p_2 - p_1$ and p_5 proves that the conditions (48) and (49) can be simultaneously satisfied only when $p_5 < 0$. We now outline this argument; in the examination of these cases we may assume that $p_1 + p_2 < 0$ (so that the fully-symmetric equilibrium exists when $r - 1 \equiv \mu > 0$) and that $a_1 > 0$. In the case $p_2 - p_1 > 0$, $p_5 > 0$ it is clear that (48) has no solutions in $\mu > 0$. In the case $p_2 - p_1 < 0$ and $p_5 > 0$, solutions of (48) exist, when $p_2 - p_1$ is sufficiently negative, but for these solutions the extra condition $\mu(p_2 - p_1) > e(p_1 + p_2)$ is never satisfied; re-arranging (48) and (49) we would simultaneously require

$$p_2 - p_1 = \frac{2a_1p_5\mu}{p_1+p_2} + \frac{2e(p_1+p_2)}{\mu} < \frac{2e(p_1+p_2)}{\mu} < 0,$$

and

$$p_2 - p_1 > \frac{e(p_1+p_2)}{\mu},$$

which is clearly not possible. When $p_5 < 0$ it is possible to satisfy (48) and (49) at the same time, and hence there exists a curve of Hopf bifurcations in the $(\mu,\ p_2-p_1)$ plane.

This line of Hopf bifurcations ends in a Takens–Bogdanov (double zero) bifurcation where the frequency of the periodic orbits created in the Hopf bifurcation tends to zero and the steady-state bifurcation coincides with the Hopf bifurcation. The location of this Takens–Bogdanov point is therefore given by the relations

$$p_1 - p_2 = \frac{2a_1p_5\mu}{p_1+p_2} = -\frac{e(p_1+p_2)}{\mu}, \tag{50}$$

which have the solution

$$\mu = \sqrt{\frac{e(p_1+p_2)^2}{2a_1|p_5|}}, \qquad p_2 - p_1 = -\sqrt{2ea_1|p_5|}. \tag{51}$$

This result is unsatisfactory in two ways. Firstly, (48) - (51) still contain the parameter e which we will later wish to exploit as a small parameter in the asymptotic analysis. Secondly, the the Takens–Bogdanov bifurcation is a codimension–2 phenomenon (generically we must vary two parameters to locate and unfold it), and this point would be made much clearer by the definition of a second parameter, in addition to μ. As the coefficients p_j are themselves functions of, for example, several cross-immunity parameters σ_i, we could select one of these σ_i as a second parameter. To keep full generality, though, we instead introduce a new parameter proportional to $p_2 - p_1$.

The relations (50) and (51) indicate the scaling required to make further analytic progress in the asymptotic limit of small e. We note that the above results hold only near $r = 1$, where we have already identified that $\Lambda_S \sim \mu \ll 1$. In addition, (48) and (50) show that, in the limit of small e, the correct balance is $\mu \sim \sqrt{e} \ll 1$. Furthermore, (51) indicates that the Takens–Bogdanov point occurs when $p_2 - p_1$ is small, in fact $\mathcal{O}(\sqrt{e})$. Hence oscillations of the kind associated with a Takens–Bogdanov point (if they occur at all) must occur near the region where the fully-symmetric and edge equilibria change stabilities, mediated by the branch of asymmetric equilibria, illustrated in figure 2.

2.4.2. Second attempt at locating the Takens–Bogdanov point, using consistent scalings

Starting from the 3×3 submatrix of the full Jacobian matrix of (40) - (47), evaluated at the fully-symmetric equilibrium, we introduce the two scaled parameters $\hat{\beta}$ and $\hat{\mu}$, defined by $p_2 = p_1 + \hat{\beta}\sqrt{e}$ and $r = 1 + \hat{\mu}\sqrt{e}$. We can then expand all variables in powers of $\sqrt{e}$ at the fully-symmetric equilibrium, in a consistent fashion: $\Lambda_S = -\hat{\mu}\sqrt{e}/p_1, U = 2\Lambda_S = -2\hat{\mu}\sqrt{e}/p_1, S_S = a_1\Lambda_S = -a_1\hat{\mu}\sqrt{e}/p_1$, $S_{12} = a_2\Lambda_S = -a_2\hat{\mu}\sqrt{e}/p_1$ and $\phi_S = 2\hat{\mu}\sqrt{e}$ to leading order. Then the 3×3 submatrix becomes (again, keeping only the leading order terms in each entry):

$$\begin{pmatrix} 0 & \frac{\hat{\mu}}{2\sqrt{e}p_1} & 0 \\ \hat{c}\sqrt{e} & -1 & \frac{p_5\hat{\mu}\sqrt{e}}{p_1} \\ a_1 & 0 & -1 \end{pmatrix} \tag{52}$$

where

$$\hat{c} = \hat{\beta} + \frac{\hat{\mu}}{p_1}\left[a_1p_6 + 2a_2(\tau_3 f_0 - \tau_5 d_1) - 2f_0 - f_0p_1\right] + \mathcal{O}(\sqrt{e}).$$

The submatrix has eigenvalues given by roots of the polynomial

$$\mathcal{P}(\lambda) = \lambda^3 + 2\lambda^2 + \lambda\left[1 - \frac{\hat{c}\hat{\mu}}{2p_1}\right] + \frac{\hat{\mu}}{2p_1}\left[-\hat{c} - \frac{a_1p_5\hat{\mu}}{p_1}\right].$$

We now have two parameters to use to locate bifurcations: $\hat{\mu}$ and $\hat{\beta}$, and the asymptotic description is consistent because the small parameter e no longer appears in $\mathcal{P}(\lambda)$. Working as before, there is a steady-state (in fact, pitchfork) bifurcation on the line

$$\hat{\beta} = -\frac{\hat{\mu}}{p_1}\left[a_1(p_6 + p_5) + 2a_2(\tau_3 f_0 - \tau_5 d_1) - f_0(p_1 + 2)\right],$$

and there is a Hopf bifurcation along the line

$$\hat{\beta} = \frac{4p_1}{\hat{\mu}} + \frac{\hat{\mu}}{p_1}\left[a_1(p_5 - p_6) - 2a_2(\tau_3 f_0 - \tau_5 d_1) + f_0(p_1 + 2)\right],$$

as long as $\hat{\mu}^2 > 2p_1^2/(a_1|p_5|)$. The oscillation frequency ω of this Hopf bifurcation is given by

$$\omega^2 = 1 - \frac{\hat{c}\hat{\mu}}{2p_1}$$

which shows that ω is order 1 in this scaling, and not larger. Because ω is order 1 we do not need to rescale time in the analysis. The epidemiological consequences of this, for interpretation of the timescale of these oscillations, is discussed in section 4. As before we find that the Hopf bifurcation can exist only when $p_5 < 0$. The Takens–Bogdanov point is the point of intersection of these two curves in the $(\hat{\mu}, \hat{\beta})$ plane:

$$(\hat{\mu}_{TB}, \hat{\beta}_{TB}) = \left(\frac{\sqrt{2}|p_1|}{\sqrt{a_1|p_5|}}, \frac{\sqrt{2}\left[a_1(p_5 + p_6) + 2a_2(\tau_3 f_0 - \tau_5 d_1) - f_0(p_1 + 2)\right]}{\sqrt{a_1|p_5|}}\right).$$

2.5. The Takens–Bogdanov bifurcation

In this section we implement the scalings deduced previously, and reduce the dynamics, via rescalings and a centre manifold reduction, to a 2D set of ODEs which describe the Takens–Bogdanov bifurcation completely. This enables us to provide a complete description of the nonlinear behaviour of the original model, in the limit of small e. Numerical investigations of the various specific models we consider below indicate that the behaviour of the models over a much larger range of e is exactly that of a Takens–Bogdanov bifurcation. In this way we are able to explain, in a unified way, the effects observed by a variety of authors.

The analysis proceeds in three stages, details of which are relegated to the Appendix. After the reduction has been carried out, the resulting equations are of the form

$$\dot{x} = y, \tag{53}$$

$$\dot{y} = -\lambda x + \kappa y + Px^3 + Qx^2 y, \tag{54}$$

where λ and κ are the 'unfolding' or bifurcation parameters and the coefficients P and Q determine the behaviour near the Takens–Bogdanov point. In this new coordinate system, the codimension–2 Takens–Bogdanov point occurs at $\lambda = \kappa = 0$

where the linearisation about the equilibrium at the origin has two zero eigenvalues. The equilibrium point $x = y = 0$ corresponds to the fully-symmetric equilibrium in the original generalised model. The non-trivial equilibria at $(x, y) = (\pm\sqrt{\lambda/P}, 0)$ correspond to the asymmetric equilibria discussed in section 2.3. The trivial solution undergoes a pitchfork bifurcation when λ passes through zero and a Hopf bifurcation when $\kappa = 0$ and $\lambda > 0$. The other bifurcations that exist depend only on the signs of P and Q; after rescaling (53) - (54) we can set both P and Q to take the values ± 1. The equations (53) - (54) are symmetric under a change of sign of both variables x and y jointly. This special feature is inherited from the symmetry of the original model (27) - (38) under the permutation ρ of the indices of the Λ_j and S_J.

For the generalised model, where we have no further information about the coefficients, the analysis presented in the Appendix shows that Q is determined by the coefficient of the nonlinear term $\Lambda_D^2\phi_D$ in the $\dot{\Lambda}_D$ equation in (66), and P is determined by the $\mathcal{O}(\sqrt{e})$ terms $B\Lambda_D^2 S_D + C\Lambda_D S_D^2$ in the $\dot{\phi}_D$ equation in (66). Using the fact that (to leading order) $p_1 = p_2$ at the Takens–Bogdanov point, we obtain

$$\begin{aligned} P &= -\frac{1}{16p_1}\left[p_4(a_1p_6 - 2f_0 + p_5) - p_6p_1(2b_0 - c_0)\right. \\ &\quad \left. +2(a_2p_4 + b_1p_1)(\tau_3 f_0 - \tau_5 d_1)\right], \\ Q &= \frac{p_1}{16\mu}. \end{aligned}$$

It is clear immediately that Q is negative in all the cases of interest in this paper, since both p_1 and p_2 are negative, as discussed above in section 2.4. However, P could take either sign. The dynamics near a $\mathbb{Z}_2$-symmetric Takens–Bogdanov point are well-known; for example see section 7.3 of the textbook by Guckenheimer & Holmes [12]. We now present a short summary of the dynamics in the two cases of interest for this paper; (i) $P > 0$ and $Q < 0$ and (ii) $P, Q < 0$. For completeness, we remark that the dynamics in the cases where $Q > 0$ can be obtained from the cases where $Q < 0$ by applying a time reversal symmetry to (53) - (54); this has no implication for the analysis in the remainder of the paper.

2.5.1. Case (i): $P > 0$ and $Q < 0$

The dynamics in this case is summarised in figure 3; figure 3(a) illustrates the regions of the (κ, λ) plane which contain qualitatively different dynamics. Figure 3(b) sketches the bifurcations which occur as (κ, λ) are varied along the two paths ℓ_1 and ℓ_2 indicated in figure 3(a).

For these values of P and Q the pitchfork bifurcation from the trivial solution (which occurs on the line labelled *pf*) is subcritical and creates two unstable (saddle-type) equilibria. The Hopf bifurcation (along the half-line labelled H) is supercritical and creates a stable periodic orbit. By rescaling x, y, λ and κ we can set $P = 1$ and $Q = -1$; only the signs of P and Q are important as far as the qualitative behaviour is concerned. From computing the linearisation at the non-trivial equilibria $(\pm\sqrt{\lambda}, 0)$ we find that they undergo no bifurcations other than the pitchfork bifurcation in which they are created when $\lambda = 0$. The stable periodic

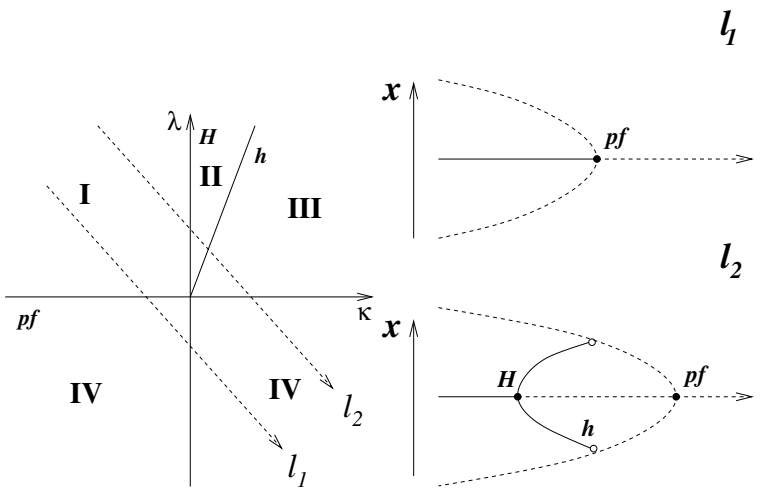

Fig. 3. Case (i): $P > 0$ and $Q < 0$. (a) Regions I - IV of the (κ, λ) plane contain qualitatively different dynamical behaviour. (b) Bifurcation diagrams seen as the two lines ℓ_1 and ℓ_2 in (a) are traversed from upper left to lower right. Local (global) bifurcations are indicated by solid (open) circles.

orbit created in the Hopf bifurcation is destroyed in a global bifurcation (labelled *h*) when it collides with the two non-trivial equilibria. This event occurs near the line $\kappa = \lambda/5$ when κ and λ are small. The location of this global bifurcation can be estimated by using a rescaled version of (53) - (54) - see [12] for details. These bifurcations are summarised in figure 3; in region IV the only equilibrium is the saddle-point at the origin, in region I the origin is a stable node or focus and the two non-trivial equilibria exist and are saddle-points. In region II the origin has now become an unstable focus, surrounded by a stable periodic orbit. In region III the origin is still an unstable focus or node but the periodic orbit has disappeared after the global bifurcation with the non-trivial equilibria.

2.5.2. Case (ii): $P < 0$ and $Q < 0$

This case is summarised in figure 4. Both the Hopf bifurcation, which occurs along the half-line *H1*, and the pitchfork bifurcation, which occurs on the line labelled *pf*, from the trivial solution at the origin are supercritical in this case. In region I the only equilibrium that exists is the origin, and it is stable. In region II a stable periodic orbit surrounds the (now unstable) origin, and in region III a further two non-trivial equilibria are also contained within it. On the line H2, where $\kappa + \lambda = 0$ and $\lambda < 0$ (after rescaling $P = Q = -1$ as in case (i)) the non-trivial equilibria undergo a subcritical Hopf bifurcation, creating an additional pair of (unstable) periodic orbits in region IV, again inside the encircling stable one. On the line marked *gluing* at $\kappa + 4\lambda/5 \approx 0$ these two unstable periodic orbits collide with the equilibrium at the origin and 'glue' to form one larger unstable periodic orbit. This exists in region V. As we approach the line labelled *sn* the large stable and unstable periodic orbits collide and disappear in a saddle-node bifurcation of periodic orbits. Region VI

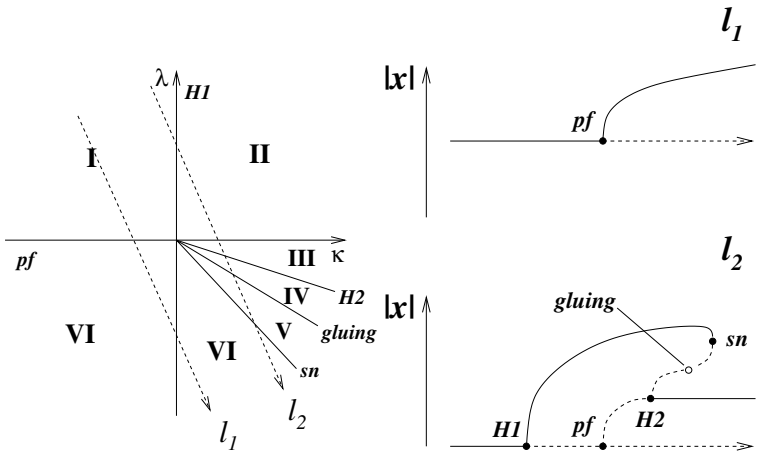

Fig. 4. Case (ii): $P < 0$ and $Q < 0$. (a) Regions I - VI of the (κ, λ) plane containing qualitatively different dynamical behaviour. (b) Bifurcation diagrams seen as the two lines ℓ_1 and ℓ_2 in (a) are traversed from upper left to lower right. Local (global) bifurcations are indicated by solid (open) circles. The saddle–node bifurcation *sn* and the gluing bifurcation are explained in the text.

therefore contains no periodic orbits and here the origin is a saddle-point and the two non-trivial equilibria exist and are stable. The dynamics in this case are markedly different from case (i); in particular we note that there is a Hopf bifurcation from the non-trivial equilibria (which correspond to asymmetric equilibria in the full models). In both cases it is possible to find a region of stable oscillations.

In the next section we apply this analysis to specific models and deduce bifurcation diagrams describing their dynamics.

3. The dynamics of four specific models

We now use the reduction outlined in the previous section to describe the dynamics of four specific models which are special cases of the generalised model. These four models are the status-based model of Gog & Swinton [8], the history-based model of Andreasen et al. [1], the reduced transmissibility approach of Ferguson & Andreasen [6] and Gupta et al. [13] and the model of Gomes et al. [10]. We will refer to them by the abbreviations GS, ALL, FA or GFA and GMN respectively. Each has particular features which we discuss in turn.

The reduced transmissibility framework investigated by FA is history-based and (for the particular strain structure used by FA, using the 'minimum' formulation of cross-immunity) it has the appealing feature that the $n + 2^n$ equations needed for describing the dynamics of n strains can be reduced by defining new linear combinations of the variables, into only $3n$ equations. This reduction is not possible for the ALL and GS models, which have reduced susceptibility. The ALL model is history-based, and was compared with the FA / GFA model by Ferguson & Andreasen [6] using numerical simulations. The GS model differs from the ALL model in that it is status-based, not history-based. This distinction is discussed in

Table 1. Choices of the generalised model coefficients corresponding to each of the specific models. Part 1. Note that for consistency, and to emphasise similarities between the GS and ALL models, σ has been replaced by $1-\sigma$ in the GS column. The notation in the ALL column is exactly as in Andreasen et al. [1]. The column headed '6-parameter' gives coefficients for the more general version of the GS model, which we discuss in section 3.3.

	GS	ALL	FA / GFA	6-parameter
τ_1	1	σ	τ_1	$\gamma_1 t_1$
τ_2	1	1	τ_2	$\gamma_2 t_2$
τ_3	1	σ	τ_3	$\gamma_1\gamma_2 t_1 t_2$
τ_4	1	σ	τ_4	$\gamma_1^2 t_1^2$
τ_5	1	σ	τ_5	$\gamma_1^2\gamma_2 t_1^2 t_2$
a_1	σ^2	1	1	$\sigma_1^2\sigma_2$
a_2	$\sigma(1-\sigma)$	0	0	$\sigma_1(1-\sigma_1)\sigma_2$
a_3	0	0	0	$(1-\sigma_2)\sigma_1^2$
a_4	0	0	0	$\sigma_1(1-\sigma_1)(1-\sigma_2)$
a_5	$(1-\sigma)^2$	0	0	$(1-\sigma_1)^2\sigma_2$
b_0	1	σ	1	γ_1
b_1	σ	σ	1	$\gamma_1\sigma_1\sigma_2$
b_2	$1-\sigma$	0	0	$\gamma_1(1-\sigma_1)\sigma_2$
b_3	0	0	0	$\gamma_1\sigma_1(1-\sigma_2)$
c_0	1	1	1	γ_2
c_1	σ^2	1	1	$\gamma_2\sigma_1^2$
c_2	$\sigma(1-\sigma)$	0	0	$\gamma_2\sigma_1(1-\sigma_1)$
d_0	1	σ	1	$\gamma_1\gamma_2$
d_1	σ	σ	1	$\gamma_1\gamma_2\sigma_1$
e_0	1	σ	1	γ_1^2
e_1	1	σ	1	$\gamma_1^2\sigma_2$
f_0	1	σ	1	$\gamma_1^2\gamma_2$

section 1. A further difference is that the ALL model uses the 'minimum' version of cross-immunity and the GS model uses the product version. Gomes et al. [10] discuss both the 'product' and 'minimum' versions of their model, which also describes cross-immunity through a complicated two-parameter function. They also discuss the dynamics of more than four strains which is beyond the scope of this paper; we comment only on the GMN model in the case $n = 4$.

Tables 1 and 2 give the values of the model coefficients for each of these specific models. Tables 3 and 4 give the values of the coefficients $p_1, \ldots, p_7$ computed directly from these coefficients.

3.1. Reduced transmissibility models

In these models cross-immunity only affects the transmission coefficients in the expressions for the effective susceptible proportions of the population θ_j. Such models, with four strains, have been analysed by Gupta et al. [13] and also by Ferguson & Andreasen [6]; in these papers they are referred to as '2 locus, 2 allele' models. In the latter paper this model is compared with a reduced-susceptibility model and the dynamics are found to be very similar. Both papers comment on the

Table 2. Choices of the generalised model coefficients corresponding to each of the specific models. Part 2. The column headed GMN I lists coefficients for the 'product' cross-immunity version of the model of Gomes et al. [10], correct to only $\mathcal{O}(q)$, due to the complicated structure of the GMN model. Column GMN II similarly lists coefficients for the minimum version.

	GMN I	GMN II
τ_1	$\sigma_{\max}(1+q)/2$	$\sigma_{\max}(1+q)/2$
τ_2	$\sigma_{\max}$	$\sigma_{\max}$
τ_3	$\sigma_{\max}^2(1+q)/2$	$\sigma_{\max}(1+q)/2$
τ_4	$\sigma_{\max}^2(1+2q)/4$	$\sigma_{\max}(1+q)/2$
τ_5	$\sigma_{\max}^3(1+2q)/4$	$\sigma_{\max}(1+q)/2$
a_1	1	1
a_2	0	0
a_3	0	0
a_4	0	0
a_5	0	0
b_0	$\sigma_{\max}(1+q)/2$	$\sigma_{\max}(1+q)/2$
b_1	$\sigma_{\max}(1+q)/2$	$\sigma_{\max}(1+q)/2$
b_2	0	0
b_3	0	0
c_0	$\sigma_{\max}$	$\sigma_{\max}$
c_1	$\sigma_{\max}$	$\sigma_{\max}$
c_2	0	0
d_0	$\sigma_{\max}^2(1+q)/2$	$\sigma_{\max}(1+q)/2$
d_1	$\sigma_{\max}^2(1+q)/2$	$\sigma_{\max}(1+q)/2$
e_0	$\sigma_{\max}^2(1+2q)/4$	$\sigma_{\max}(1+q)/2$
e_1	$\sigma_{\max}^2(1+2q)/4$	$\sigma_{\max}(1+q)/2$
f_0	$\sigma_{\max}^3(1+2q)/4$	$\sigma_{\max}(1+q)/2$

Table 3. Values of the coefficients $p_1, \ldots, p_7$ for each of the specific models. Part 1. These values are computed directly from table 1.

	GS	ALL	FA / GFA
p_1	-1	-1	$\tau_2 - 2$
p_2	$2(\sigma-1)$	$2(\sigma-1)$	$2(\tau_1-1)$
p_3	$2(\sigma-1)$	$(\sigma-1)(\sigma+\frac{1}{2})$	$2\tau_3+\tau_4-2\tau_1-\tau_2$
p_4	0	$(1-\sigma)(\sigma-\frac{1}{2})$	$\tau_2+\tau_4-2\tau_3$
p_5	0	$(\sigma-1)(\sigma+\frac{1}{2})$	$2\tau_1-\tau_2-\tau_4$
p_6	$2(1-\sigma)$	$(1-\sigma)(\sigma-\frac{1}{2})$	$\tau_2-\tau_4$
p_7	$\sigma-1$	0	$\tau_5-\tau_3$

existence of a range of values for the cross-immunity parameter (labelled γ and c respectively in the two papers) where the stable dynamics is oscillatory or chaotic. For values of the parameter above and below the boundaries of this region, the long-time behaviour of the system is steady, and in one region tends towards the fully-symmetric equilibrium where all strains invade the host population equally

Table 4. Values of the coefficients $p_1, \ldots, p_7$ for each of the specific models. Part 2. Due to the complicated structure of the GMN model, we have computed these coefficients correct to $\mathcal{O}(q)$, using the values computed in table 2.

	GMN I	GMN II
p_1	$\sigma_{\max} - 2$	$\sigma_{\max} - 2$
p_2	$\sigma_{\max} - 2 + q\sigma_{\max}$	$\sigma_{\max} - 2 + q\sigma_{\max}$
p_3	$(\sigma_{\max}^4 + 3\sigma_{\max}^3 - 8\sigma_{\max}^2)/4$ $+q(5\sigma_{\max}^4 + 12\sigma_{\max}^3 - 16\sigma_{\max}^2)/8$	$-\sigma_{\max}^2/2 + q\sigma_{\max}^2/4$
p_4	$-\sigma_{\max}^3/4 - q\sigma_{\max}^2(\sigma_{\max} + 2)^2/8$	$q\sigma_{\max}^2/4$
p_5	$-\sigma_{\max}^3/4 - q\sigma_{\max}^2(\sigma_{\max} + 2)^2/8$	$-\sigma_{\max}^2/2 + q\sigma_{\max}^2/4$
p_6	$\sigma_{\max}^3(\sigma_{\max} - 1)/4 + q\sigma_{\max}^3(5\sigma_{\max} - 4)/8$	$q\sigma_{\max}^2/4$
p_7	$\sigma_{\max}^4(\sigma_{\max} - 2)/4 + q\sigma_{\max}^4(3\sigma_{\max} - 4)/4$	0

(denoted by the label 'no strain structure' in Gupta et al. [13]), while in the other region the edge equilibrium is stable, labelled as the 'dominance of one discordant set' (of strains). In the GFA paper e was varied between 0.1 and 10^{-5} but the basic reproduction ratio r was kept fixed at $r = 4$. The asymptotic analysis of sections 2.3 – 2.5 is only valid for r close to 1, and hence more bifurcations may take place as r is increased. The results of section 2.3 demonstrate that quasiperiodic oscillations and chaotic dynamics do not exist arbitrarily close to $r = 1$, yet they are observed in the numerical simulations of [13] and [6]. A natural conclusion is that the periodic oscillations, which have been shown to occur near $r = 1$, undergo further Hopf bifurcations as r increases, producing quasiperiodic oscillations, and this becomes a chaotic attractor at higher r. Similar results have been obtained by Gomes et al. [10], and we comment on their model in section 3.4 below.

Of more interest is the observation by Ferguson & Andreasen [6] that the reduced-transmissibility and reduced-susceptibility versions of their model produce nearly identical numerical results (at least for the values of r at which they performed numerical integrations of their models). The reduced-susceptibility version is, in fact, the same model as that studied by [1] (but without the 'removal of infecteds' complication), and analysed in more detail in section 3.2. We can apply the general analysis of the previous section to analyse these two models by calculating the coefficients $p_1, \ldots, p_7$ for the two models (shown in table 3) and computing the resulting values for P and Q which determine the structure of the Takens–Bogdanov bifurcation. For the reduced transmissibility model we agree with the observation of [6] that the asymmetric equilibria are stable when they exist. This is not the case for the reduced susceptibility version of the model. In the remainder of this section we analyse the reduced-transmissibility model in detail.

Since we are considering infections where previous exposure only heightens the immune response to subsequent infections (which excludes a few diseases where the reverse is true, such as dengue fever), we take $\tau_j \leq 1$. It is also reasonable to suppose that $\tau_2 \geq \tau_1 \geq \tau_3 \geq \tau_4 \geq \tau_5$, as we are assuming that cross-immunity to adjacent strains is greater than that given to the opposite strain, see section 2.1. In passing we note that to satisfy the assumption (39) made in the reduction calculation we need to take $\tau_4 = \tau_5$ in addition. The relevant coefficient values for this

model are given in the column headed 'FA / GFA' of tables 1 and 3. For this model both the fully-symmetric and edge equilibria exist in $r > 1$ since both $p_1 = \tau_2 - 2$ and $p_1 + p_2 = 2\tau_1 + \tau_2 - 4$ are negative. From the earlier results of section 2.3, the degeneracy in the bifurcation at $r = 1$ occurs when $p_1 = p_2$, i.e. when $2\tau_1 = \tau_2$. In the scaled low e limit, the Takens–Bogdanov point occurs at

$$\hat{\mu} = (2 - \tau_2)\sqrt{\frac{2}{\tau_4}},$$
$$\hat{\beta} = -2\sqrt{2\tau_4},$$

and the pitchfork and Hopf bifurcations, respectively, occur along the lines

$$\hat{\beta} = \frac{-2\tau_4\hat{\mu}}{2 - \tau_2},$$

and

$$\hat{\beta} = 4(\tau_2 - 2)/\hat{\mu}, \qquad \text{when} \qquad \mu^2 > 2(\tau_2 - 2)^2/\tau_4.$$

Computing the values of P and Q at the Takens–Bogdanov point (when $2\tau_1 = \tau_2$) and setting $\tau_4 = \tau_5$, to satisfy (39), we obtain

$$P = \frac{\tau_4(2\tau_3 - \tau_2 - \tau_4)}{8(2 - \tau_2)},$$
$$Q = \frac{2(\tau_2 - 2)}{32},$$

which indicates that P may in theory take either sign, but $Q < 0$.

Specialising to the simpler case considered by Ferguson & Andreasen [6] where $\tau_1 = \tau_3 = \tau_4 = \tau_5 = c < 1$ and $\tau_2 = 1$ we find that the bifurcation from $r = 1$ becomes degenerate at $c = 1/2$. The Takens–Bogdanov bifurcation then occurs close to the point $(r, c) = (1, 1/2)$, and in the scaled co-ordinates it is located at $(\hat{\mu}, \hat{\beta}) = (2, -2)$. The equation for the pitchfork bifurcation line becomes $\hat{\beta} = -\hat{\mu}$ and the Hopf bifurcation is located at $\hat{\beta} = -4/\hat{\mu}$ when $\hat{\mu} > 2$.

Since both P and Q are nonzero when evaluated at the degenerate point $c = 1/2$ (in fact, $P = -1/32$ and $Q = -1/16$) the reduction calculation of section 2.5 produces valid results and demonstrates that the dynamics near the fully-symmetric equilibrium, near $r = 1$, are as sketched in figure 4. The complete (qualitative) bifurcation diagram is then obtained by combining this with figure 2(a) and (c) to produce figure 5.

3.2. The history-based model of Andreasen et al. [1]

The model analysed by Andreasen et al. [1] (referred to as the ALL model) is history-based, which means that the susceptible host population is divided into classes S_J labelled by the strains that they have previously been infected by. In this way the host population is implicitly assumed to be homogeneous in its response to new infections. The model described in [1] includes 'removal of infecteds'; the

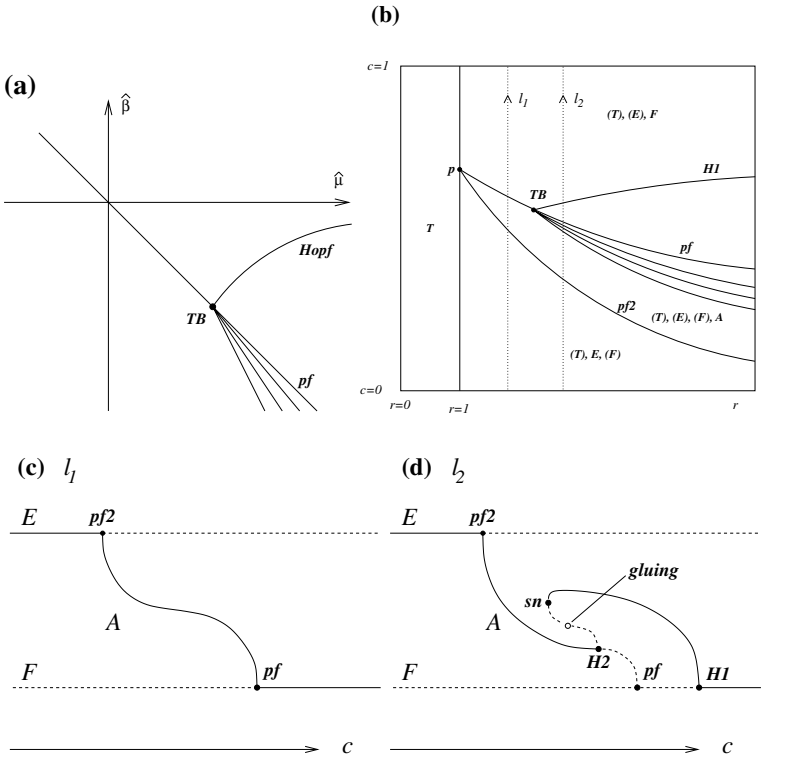

Fig. 5. The dynamics of the reduced transmissibility model. (a) The local behaviour near the scaled Takens–Bogdanov point $(\hat{\mu}, \hat{\beta}) = (2, -2)$. (b) Sketch of the (r, c) plane for small e indicating the position p of the degenerate bifurcation at $r = 1$, the Takens–Bogdanov point TB and the bifurcation lines extending from them. The local behaviour described by (a) organises the bifurcation lines near $r = 1, c = 1/2$. (c) The bifurcation sequence as c is increased for $r < r_{TB}$. The labels E, F and A refer to the edge equilibrium, the fully-symmetric equilibrium and the asymmetric equilibrium respectively. Compare with the bifurcation sequence along ℓ_1 in figure 4(b). (d) The bifurcation sequence as c is increased for $r > r_{TB}$. Compare with the bifurcation sequence along ℓ_2 in figure 4(b).

host population is not just the sum of all the S_J categories of susceptible hosts, but we also keep track of the number of infected hosts in a set of classes I^i_J: I^i_J is the number of individuals who are currently infected with strain i and who have previously recovered from infections with the strains in the set J. The version of the model that we analyse here, and the one used later by Ferguson & Andreasen [6], ignores the dynamics of these classes I^i_J. Since there are $n2^{n-1}$ of these classes, for an n-strain model, inclusion enlarges the dimension of the ODEs substantially (in the case $n = 4$ from 20 ODEs without infectious classes to 52 ODEs with them). From our viewpoint, this 'removal of infecteds' is just an extra complication that we wish to remove, as discussed in section 2.1. Moreover, in the limit of small e our analysis will not be affected by the inclusion or omission of the I^i_J variables because they can be slaved to the S_J variables and hence eliminated.

The choices of the coefficients in the generalised model (5) - (11) which give the ALL model are given in column 2 of table 1. There is one cross-immunity parameter σ (labelled c in [6]), which satisfies $0 \leq \sigma \leq 1$.

Just as for the reduced transmissibility model, as r increases through 1 there are two distinct bifurcating branches of equilibria. They take the form $\Lambda_1 = \Lambda_2 \neq 0$ (the 'fully symmetric' equilibrium) and $\Lambda_1 \neq 0$, $\Lambda_2 = 0$ (an 'edge' equilibrium). We do not distinguish the equilibria $\Lambda_1 \neq 0$, $\Lambda_2 = 0$ and $\Lambda_1 = 0$, $\Lambda_2 \neq 0$ since they are related by the permutation symmetry of the model. Exactly one of these types of equilibrium is stable if both branch supercritically. When $\sigma = 1/2$, there is a degeneracy in the equations, similar to the degeneracy noted for the reduced transmissibility model when $2\tau_1 = \tau_2$ (or $c = 1/2$). This is because $p_1 = p_2$ when $\sigma = 1/2$. Exactly as before, higher-order terms are required to describe the steady-state behaviour near $r = 1$, and asymmetric equilibria exist close to $\sigma = 1/2$.

Unlike the reduced transmissibility model, however, when the higher-order terms are computed we find that the asymmetric branch is *unstable*, and in the region of existence of the asymmetric equilibria, both other branches stably co-exist. The local dynamics near the Takens–Bogdanov point are illustrated by figure 3, not by figure 4. Figure 6 illustrates, exactly as figure 5 did for the reduced transmissibility model, how the bifurcations near the Takens–Bogdanov point fit into the (r, σ) plane.

Away from $r = 1$, the region of existence of the asymmetric equilibria can be found numerically by locating the pitchfork bifurcations from each of the fully-symmetric and edge equilibria. These are shown as the curved solid lines on figure 7(a); the asymmetric equilibria exist in the wedge-shaped region between the lines. Figure 7 was computed numerically using the bifurcation and continuation package AUTO [5], and should be compared with figure 2 in [1]; in that paper the pitchfork bifurcations bounding the region of existence of the asymmetric equilibria were not shown. The closeness of the two lines of pitchfork bifurcations may well be the reason that Andreasen et al did not discern them.

As is the case for all the other models discussed in this paper, none of the steady-state dynamics discussed above depends on the value of the coefficient e. For e low enough (numerical investigations show that the range is at least $e \leq 0.08$) there is a Takens–Bogdanov bifurcation on the line of pitchfork bifurcations from the fully-symmetric equilibrium. Near this bifurcation point we expect that the dynamics can be described by the two-dimensional set of ODEs (53) - (54) for some choice of the coefficients P and Q. In particular we would like to explain the location of the line of supercritical Hopf bifurcations from the fully-symmetric equilibrium; this was computed by ALL, and meets the line of pitchfork bifurcations at the Takens–Bogdanov point. A typical oscillatory solution is shown in figure 8: in contrast to the sketches in figure 2 of [1], the periodic orbit is in the shape of a figure-of-eight when projected into the (Λ_1, Λ_2) plane.

For fixed e and r, and decreasing σ, the oscillation period grows rapidly and the periodic orbit spends increasing amounts of time near to the two edge equilibria $\Lambda_1 \neq 0$, $\Lambda_2 = 0$ and $\Lambda_1 = 0$, $\Lambda_2 \neq 0$ until it is destroyed in a homoclinic bifurcation. Near the Takens-Bogdanov point numerical investigations show that the

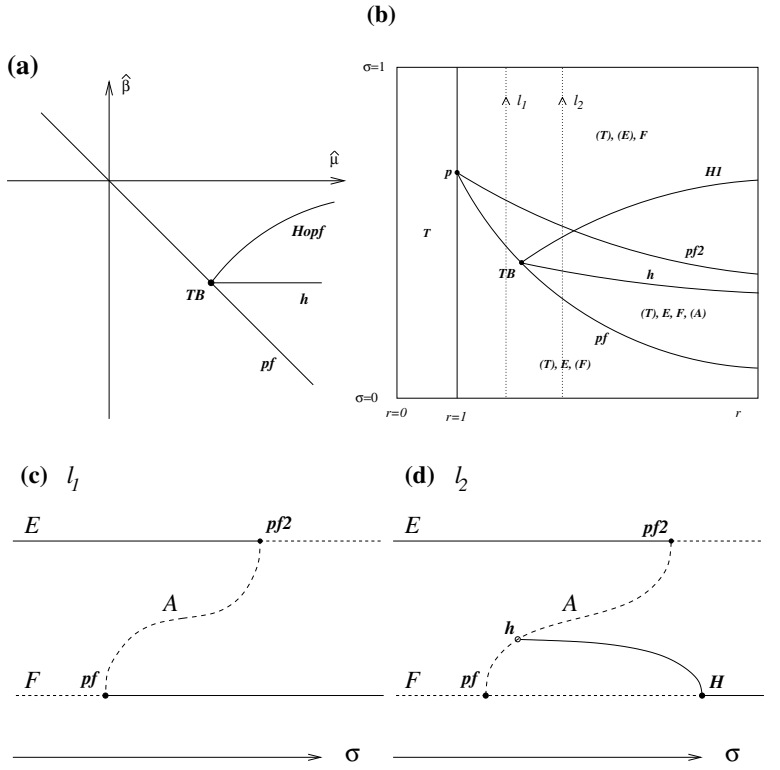

Fig. 6. Qualitative dynamics of the ALL model. (a) The local behaviour near the scaled Takens–Bogdanov point $(\hat{\mu}, \hat{\beta}) = (2, -2)$. (b) Sketch of the (r, σ) plane for small e indicating the position p of the degenerate bifurcation at $r = 1$, the Takens–Bogdanov point TB and the bifurcation lines extending from them. The local behaviour described by (a) organises the bifurcation lines near $r = 1$, $\sigma = 1/2$. (c) The bifurcation sequence as c is increased for $r < r_{TB}$. The labels E, F and A refer to the edge equilibrium, the fully-symmetric equilibrium and the asymmetric equilibrium respectively. This should be compared with the bifurcation sequence along ℓ_1 in figure 3(b). (d) The bifurcation sequence as c is increased for $r > r_{TB}$. This should be compared with the bifurcation sequence along ℓ_2 in figure 3(b).

periodic orbit collides in a homoclinic bifurcation with the (unstable) asymmetric solutions which exist between the two solid curved lines of pitchfork bifurcations. This curve of homoclinic bifurcations is indicated by the dashed line in figure 7(b).

This is the sequence of events that we would expect in case (i) of the Takens–Bogdanov bifurcation, where $P > 0$ and $Q < 0$. Applying the analysis of the generalised model directly, we find that the degenerate point where $p_1 = p_2$ occurs at $\sigma = 1/2$. Evaluating the coefficients $p_1, \ldots, p_7$ at $\sigma = 1/2$, we find that the pitchfork bifurcation from the fully-symmetric equilibrium occurs at $\hat{\beta} + \hat{\mu} = 0$ in the scaled variables of section 2.3. Similarly the Hopf bifurcation occurs when $\hat{\beta} = -4/\hat{\mu}$ as long as $\hat{\mu} > 2$, and these lines meet at the Takens–Bogdanov point $(\hat{\mu}, \hat{\beta}) = (2, -2)$ just as for the reduced transmissibility model. However,

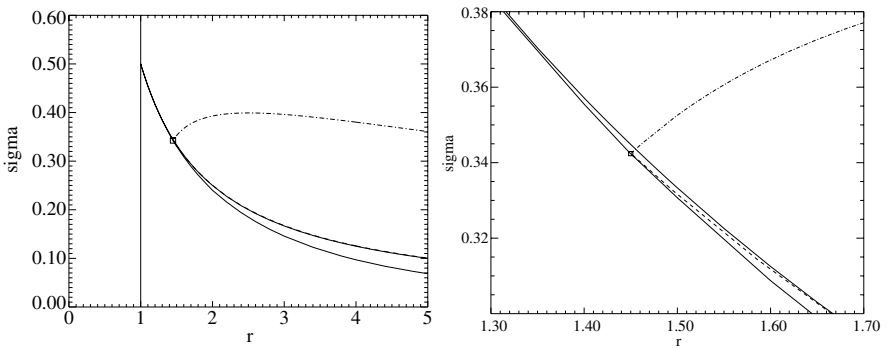

Fig. 7. Bifurcation structure in the ALL model, computed using AUTO. (a) Regions of the (r, σ) plane showing qualitatively distinct behaviour, separated by numerically computed bifurcation curves. When $r < 1$ the trivial solution is stable. For $r > 1$ the fully-symmetric and edge equilibria exist and at least one is stable. The fully-symmetric equilibrium is stable above the dot-dashed line which denotes the Hopf bifurcation. The edge equilibrium is stable below the higher curved solid line. Oscillations exist between the dot-dashed line and the long-dashed line which lies very close to the upper solid line. (b) Enlargement of (a) showing the bifurcation structure near the Takens-Bogdanov (TB) point for $e = 0.02$. The □ symbol indicates the TB point, where the Hopf bifurcation (dot-dashed line) and the pitchfork bifurcation (lower solid line) meet. The line of homoclinic bifurcations where the periodic orbit is destroyed is indicated by the long dashed line and asymptotes to the upper solid line at large r.

evaluating the Takens–Bogdanov coefficients P and Q results in $P = 0$ and $Q = -1/16$ which means we cannot decide which of case (i) or (ii) occurs using the leading-order results of the calculation described in the Appendix. Repeating this calculation, and keeping higher-order terms, gives the following nonlinear terms in the reduced description for the Λ_D, ϕ_D, S_D variables (compare with (66)):

$$\dot{\Lambda}_D = \frac{1}{4}\Lambda_D^2\phi_D + \sqrt{e}\Lambda_D\phi_D\left(\frac{1}{8}\phi_D + \frac{1}{2}\Lambda_D\right) + \mathcal{O}(e)\phi_D, \tag{55}$$

$$\dot{\phi}_D = -\frac{\sqrt{e}}{8}\Lambda_D\phi_D(\Lambda_D + S_D) - \frac{e}{8}\Lambda_D S_D(2\Lambda_D + S_D) + \mathcal{O}(e)\phi_D, \tag{56}$$

$$\dot{S}_D = \mathcal{O}(e)\phi_D + \frac{e\sqrt{e}}{2}\Lambda_D^2 S_D + \mathcal{O}(e\sqrt{e})\phi_D. \tag{57}$$

In (55) - (57) we have included all the terms at $\mathcal{O}(\sqrt{e})$, the one term (in (56)) at $\mathcal{O}(e)$ which does not contain a factor of ϕ_D, and have indicated the remaining terms at $\mathcal{O}(e)$ (which all contain factors of ϕ_D) by "$\mathcal{O}(e)\phi_D$". The $\mathcal{O}(e)$ term which does not contain a factor of ϕ_D is crucial since, after the linear change of co-ordinates $(\Lambda_D, \phi_D, S_D) \to (x, y, z)$ detailed in the Appendix has been carried out, it is only this term which contributes to the coefficient of x^3 in the resulting equation for $\dot{y}$. This coefficient of x^3 is, of course, the coefficient P. The result of this calculation is $P = 3e/64$ which is positive, showing that the inclusion of these higher-order terms has broken the degeneracy in the leading-order calculation, and enables us to distinguish between the two cases.

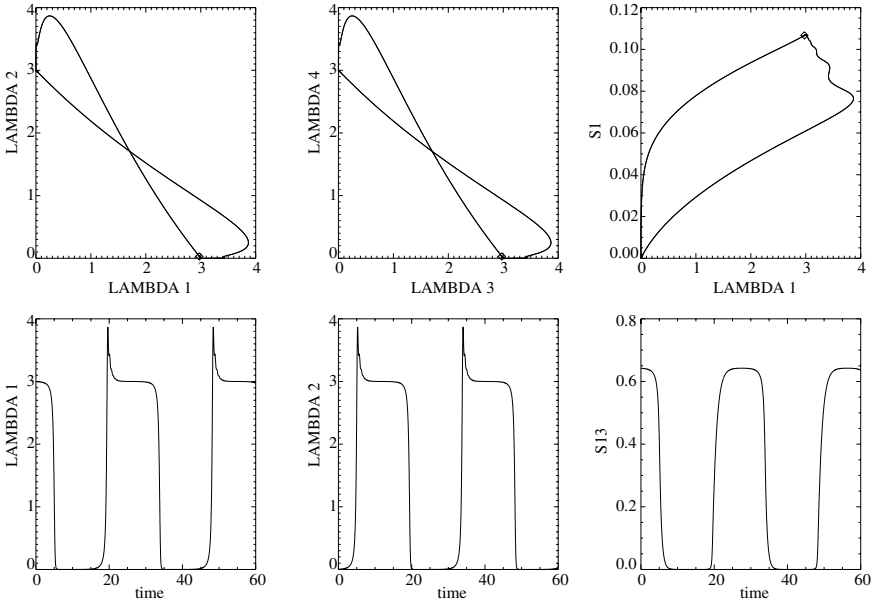

Fig. 8. An oscillatory solution for the ALL model at $r = 4.0$, $\sigma = 0.13$, $e = 0.02$, close to the heteroclinic bifurcation showing the figure-of-eight structure. The first three plots are projections into three co-ordinate planes, the final three plots show the time evolution of Λ_1, Λ_2 and S_{13}.

Further away from the Takens–Bogdanov point, the location of the homoclinic bifurcation approaches the location of the pitchfork bifurcation from the edge equilibrium from lower values of σ. This means that the pitchfork bifurcation destroying the unstable mixed-mode solutions and the homoclinic bifurcation happen almost at the same parameter values. However, the two lines do not intersect, and the homoclinic bifurcation always happens at a slightly lower value of σ (for given values of e and r) than the pitchfork bifurcation. This is because the global bifurcation always involves the asymmetric equilibria and never the edge equilibria. The edge equilibria are not involved due to the flow-invariance of the co-ordinate planes which contain them; it is not possible to have a heteroclinic trajectory between two edge equilibria since the stable manifold of each of these equilibria lies entirely within the relevant invariant co-ordinate plane.

For this model we have also investigated the range of e over which oscillatory solutions exist. It seems that e does not have to become very large before oscillatory solutions are confined to a very small region of the (r, σ) plane, and then disappear completely. For $e = 0.08$, for example, the line of Hopf bifurcations meets the pitchfork bifurcation line again at larger r, at a second Takens-Bogdanov bifurcation, so oscillatory solutions are confined to the enclosed bubble-like region. As e decreases, the first Takens-Bogdanov point moves to lower r and larger σ, and the second one moves to larger r - its exact position has not been investigated.

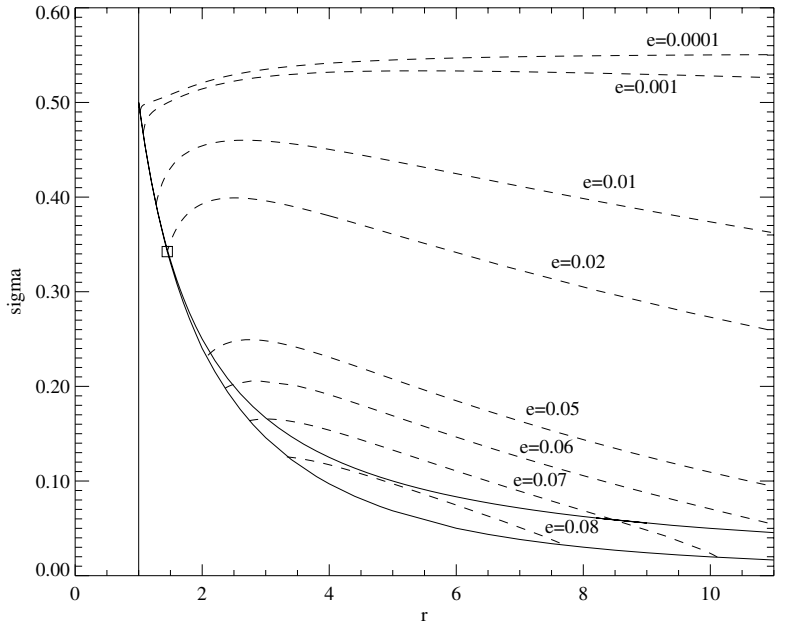

Fig. 9. Variation of the location of the Hopf bifurcation curve with e, for the ALL model. From top to bottom, the dashed lines indicate the position of the Hopf bifurcation for $e = 0.0001, 0.001, 0.01, 0.02, 0.05, 0.06, 0.07, 0.08$. For larger e note that the dashed line meets the pitchfork bifurcation line at both ends, and there are therefore two Takens-Bogdanov points. For a fixed value of e, oscillatory solutions exist in a small bubble between the dashed curve and the lower of the two solid curves.

Hence for lower e there is a larger region of stable oscillatory behaviour. As e tends to zero (i.e. the proportion of the host's lifetime for which it is infected with any strain), the Takens-Bogdanov point tends towards $r = 1$, $\sigma = 0.5$, and the Hopf bifurcation line seems to become flatter, lying closer to $\sigma = 0.5$. These numerical observations are consistent with our theoretical analysis at low e. The location of the Hopf bifurcation line for different values of e is shown in figure 9.

3.3. The status-based model of Gog & Swinton [8]

Since the values of the coefficients $p_1 = -1$ and $p_2 = 2(\sigma - 1)$ are identical for the ALL and GS models, it is no surprise that the qualitative steady-state features of the two models are the same. In both cases the fully-symmetric equilibrium is stable for large σ and the edge equilibrium is stable for small σ (note that we are using σ defined the opposite way around from [8], transforming $\sigma \to 1 - \sigma$). The location of the two pitchfork bifurcations in the GS model can be found analytically and is given in that paper. Figure 10 (computed using AUTO) compares the location of the steady bifurcation curves, and the region of existence of the asymmetric equilibria in the two models.

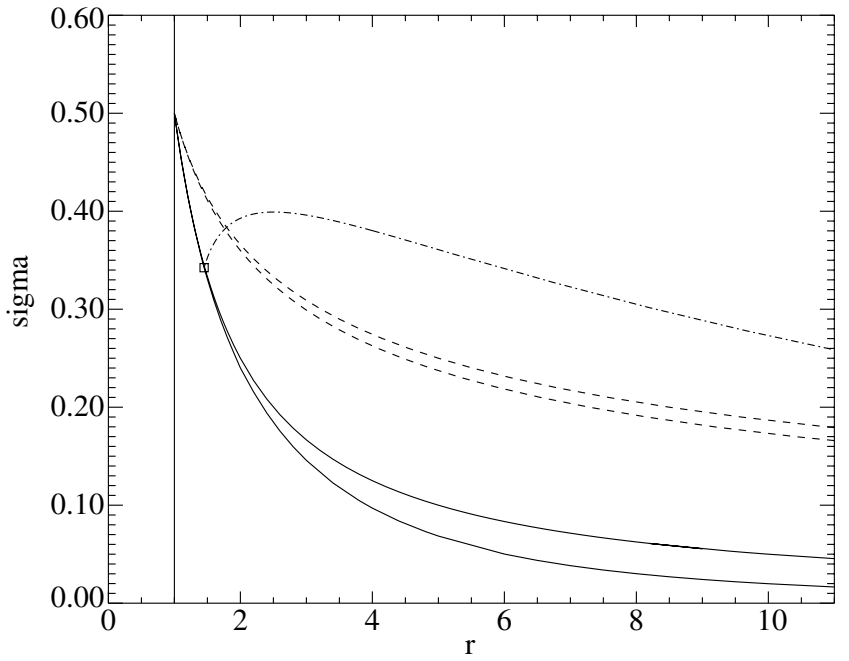

Fig. 10. Comparison of the behaviour of the ALL and GS models. The solid curves indicate the steady bifurcations from the fully-symmetric and edge equilibria for the ALL model; asymmetric equilibria exist between these lines and are unstable. The steady bifurcations from the fully-symmetric and edge equilibria in the GS model are indicated by the dashed lines; asymmetric equilibria exist (unstably) between these dashed curves. The dash-dotted curve indicates the Hopf bifurcation in the ALL model (for $e = 0.02$) and is the only curve that whose location varies with e.

However, the overall dynamics are very different: numerical investigations by Gog & Swinton [8] showed that the GS model does not contain any oscillatory dynamics near $r = 1$. We are able to explain this apparent anomaly using the analysis of section 2.4. It turns out that the coefficient p_5 is identically zero for the GS model, and so the matrix (52) has one eigenvalue -1 and the other two sum to -1 and so cannot be a purely imaginary pair. So no Hopf bifurcation from the fully-symmetric equilibrium is possible, even for very small e, and hence no Takens–Bogdanov bifurcation can exist.

There are two possible reasons for the lack of oscillations; either there is something implicit in the assumptions behind the GS model that prohibits oscillations, or the coefficients have just conspired to make p_5 identically zero. We take the view that it is the latter reason that applies here. Our reasoning is supported by the analysis of another model which is quite general, but more specific than the generalised model (5) - (11). In it we express the coefficients $a_1, \ldots, f_0$ of the generalised model (5) - (11) in terms of six parameters σ_1, σ_2, γ_1, γ_2, t_1 and t_2. σ_1 gives the probability that total immunity to the two adjacent strains is not acquired as a result of infection with one strain. σ_2 gives the similar probability

that immunity to the opposite strain is not acquired after an infection. So we naturally expect $0 \leq \sigma_1 \leq \sigma_2 \leq 1$. γ_1 gives the factor by which susceptibility to an adjacent strain is reduced after an infection, and γ_2 similarly gives the factor by which susceptibility to the opposite strain is reduced, hence $0 \leq \gamma_1 \leq \gamma_2 \leq 1$. Similarly, t_1 and t_2 give the factors by which transmissibility of an adjacent or opposite strain is reduced, respectively, so $0 \leq t_1 \leq t_2 \leq 1$. Table 1 lists the coefficients $a_1, \ldots, f_0$ in terms of these six parameters. By way of illustration, the choice $\gamma_1 = \gamma_2 = t_1 = t_2 = 1, \sigma_1 = \sigma, \sigma_2 = 1$ corresponds to the GS model.

It is now straightforward to compute p_5 in terms of these six parameters. The result is

$$p_5 = \gamma_1^3\gamma_2 t_1^2 t_2(1-\sigma_2) + \gamma_1^2\gamma_2 t_1^2(1-\sigma_1)[1-\gamma_1 t_2 + \sigma_1(1-\gamma_2 t_2)] + 2\gamma_1^2 t_1 + \gamma_1\gamma_2(t_1 - t_2) - \frac{1}{2}\gamma_2^2 t_2 + \frac{1}{2}\gamma_1^2\gamma_2^2 t_2 - \gamma_1^3\gamma_2 t_1 - \gamma_1^2\gamma_2 t_1^2. \quad (58)$$

The form of this expression means we can draw a couple of general conclusions about conditions which promote oscillations. Firstly, both the coefficient of $(1-\sigma_2)$ (the first term) and the factor in square brackets in the second term are non-negative. Hence any deviation of the parameters σ_1 and σ_2 away from 1 will tend to inhibit oscillations since this would make p_5 more positive (oscillations are only possible if $p_5 < 0$ at the point where $p_1 = p_2$, as discussed in section 2.4). When $\sigma_1 = \sigma_2 = 1$, it is possible to make p_5 negative, and so promote oscillations, by reducing the 'susceptibility' parameters γ_1 and γ_2 to below 1. More specifically, for the GS model consider decreasing τ_1 to slightly below unity while keeping the other $\tau_j = 1$. Then $p_1 = -1$ and $p_2 = 2(\tau_1\sigma^2 + \sigma - \sigma^2 - 1)$. Hence $p_1 = p_2$ when $\sigma \approx (3-\tau_1)/4$ and, at this point, $p_5 \approx 2(\tau_1 - 1)$ which is negative, allowing oscillations to occur. This is illustrated in figure 11 where $\tau_1 = 0.8$ and oscillations exist in a small region near the transition from the stable fully-symmetric equilibrium to the stable edge equilibrium.

These observations from the six-parameter model help to explain, at least mathematically, why the reduced susceptibility and reduced transmissibility models may contain oscillations, but the status-based model of Gog & Swinton [8] does not, although perturbed versions of the GS model do contain oscillations.

3.4. Comparison with the model of Gomes et al. [10]

Our final specific model is that developed by Gomes et al. [10], and investigated by these authors for four to eight disease strains. This paper focussed on comparing aspects of the dynamics for different numbers of strains, and also compared results obtained with 'product' and 'minimum' cross-immunity structures. Unlike the GS model, this is a history-based model. The GMN model differs from the ALL model in that the various cross-immunities are given as functions of two independent parameters $\sigma_{\max}$ and p, whereas ALL had only one (labelled σ). The positive parameter $\sigma_{\max}$ controls the overall range of cross-immunity available, and p controls the relative strength of local interactions as opposed to longer-range ones; distance between strains is a measure of how closely related they are antigenically, and hence how strong the cross-immunity protection is. In this formulation the case $\sigma_{\max} = 1$,

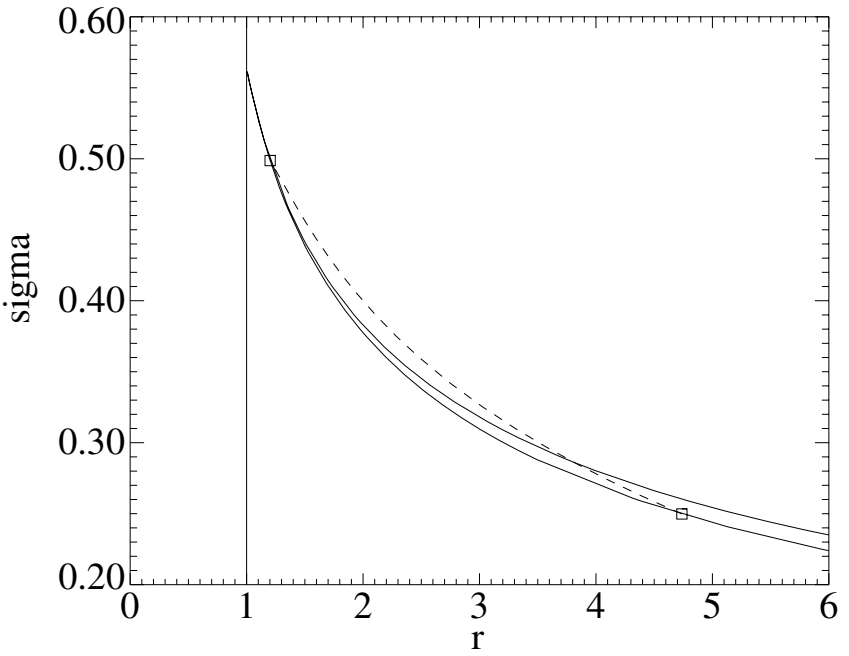

Fig. 11. Dynamics of the perturbed GS model for $e = 0.001$, with $\tau_1 = 0.8$. Oscillations exist in the thin crescent-shaped region below the dashed line and above the lower solid line, between the two squares. The fully-symmetric equilibrium is stable above the dashed line; this line gives the location of the Hopf bifurcation. The two solid lines indicate the pitchfork bifurcations, as in figure 10. Note the slightly altered axis limits compared to figure 10. The squares indicate Takens–Bogdanov points where the Hopf bifurcation curve ends, as in figure 9.

$p = 0$ corresponds exactly to the ALL model with $\sigma = 1/2$. As p decreases below 0 the local cross-immunity interactions become more strongly promoted compared to the longer-range ones. For comparison we note that $\sigma_{\max}$ varies in the opposite way to σ in the ALL model; $\sigma_{\max}$ increasing corresponds to σ decreasing.

Despite these differences, the observed dynamics of the GMN model are strikingly similar to those of the ALL model. In particular, figure 3(b) in [10] closely resembles figure 2 in [1]. By performing the asymptotic analysis at low e on the GMN model these similarities become clear: Gomes et al. assumed an infectious period of one month and a lifetime of 67 years, yielding a value for the dimensionless parameter e of $e = 0.00125$ which is certainly small enough for their results to lie within the region of validity of our asymptotic results (for the ALL model oscillations persist for e up to approximately $e = 0.08$).

In the remainder of this section we analyse the GMN model (but, of course, only in the case of four interacting strains) and demonstrate why the two figures referred to above are so similar.

Table 2 shows the values of the generalised model coefficients $a_1, \ldots, f_0$ for the GMN model, for both product cross-immunity and minimum cross-immunity. In these columns of the table we have introduced the rescaled version $q = 3\pi p/32$ of

the cross-immunity parameter p defined in [10] which controls the relative balance of local and long-range contributions to cross-immunity. We are able to simplify the formula used in [10] by considering only the region where q is small, and ignoring $\mathcal{O}(q^2)$ contributions to the coefficients $a_1, \ldots, f_0$; it is clear from figure 3(b) that the interesting region of oscillatory behaviour occurs for p, and hence q, close to zero.

From the generalised model coefficients we compute the combinations $p_1, \ldots, p_7$ and they are given in table 4. As before, these are correct to $\mathcal{O}(q)$ only. Note that $p_1 - p_2$ is $\mathcal{O}(q)$; the region of interest corresponds to small q. From the analysis of steady-state dynamics near $r = 1$ in section 2.3 we observe that the two kinds of equilibrium both exist in $r > 1$. When $q < 0$ ($p < 0$ in the GMN notation) the edge equilibrium is stable and the fully-symmetric one is unstable. When $q > 0$ the fully-symmetric one is stable and the edge equilibrium is unstable. This agrees with figure 3(b) of [10], near $r = 1$. However, since e is so small, the oscillatory dynamics occur very close to $r = 1$ in this figure. The degeneracy in the bifurcation at $r = 1$ occurs when $p_1 = p_2$, i.e. when $q = 0$, for any value of $\sigma_{\max}$.

We now analyse the product and minimum cross-immunity cases of the GMN model separately, giving results for the product cross-immunity case first. For the product cross-immunity case the additional assumption (39) holds at $\sigma_{\max} = 1$, but not otherwise. However the analytic results we obtain do not change qualitatively as $\sigma_{\max}$ is varied around 1, and the numerical results of Gomes et al. [10] give strong support to the claim that they are qualitatively correct. At the point $q = 0$ we find that $p_5 = -\sigma_{\max}^3/4 < 0$ and so the Takens–Bogdanov point exists. In the scaled co-ordinates it is located at

$$\hat{\mu} = \sqrt{\frac{8}{\sigma_{\max}^3}}(2 - \sigma_{\max}) = \hat{\mu}_{TB},$$
$$\hat{\beta} = -\sqrt{2\sigma_{\max}^3} = \hat{\beta}_{TB}.$$

The pitchfork and Hopf bifurcations from the fully-symmetric equilibrium, respectively, occur along the lines

$$\hat{\beta} = -\frac{\sigma_{\max}^3}{2(2 - \sigma_{\max})}\hat{\mu},$$

and

$$\hat{\beta} = \frac{4(\sigma_{\max} - 2)}{\hat{\mu}}, \qquad \text{when} \qquad \hat{\mu} > \hat{\mu}_{TB}.$$

Computing the values of P and Q at the Takens–Bogdanov point we obtain

$$P = \frac{\sigma_{\max}^6(4 - \sigma_{\max})}{256(2 - \sigma_{\max})},$$
$$Q = -\frac{\sigma_{\max}\sqrt{\sigma_{\max}}}{32\sqrt{2}}.$$

So, for $\sigma_{\max}$ close to 1 we find that $P > 0$ and $Q < 0$, exactly as for the ALL model. Moreover, the equations for the pitchfork and Hopf bifurcation lines when $\sigma_{\max} \approx 1$ are very similar to those obtained for the ALL model; the Hopf bifurcation is identical and the pitchfork bifurcation line differs only by a factor of $1/2$. At least for small e, then, the dynamics of the two models are the same.

For the minimum formulation of cross-immunity, the details of the analysis of the GMN model are even more similar to those for the ALL model. The coefficients $p_1, \ldots, p_7$ are given in table 4, and we note that $p_5 = -\sigma_{\max}^2/2 + q\sigma_{\max}^2/4 + \mathcal{O}(q^2)$ is again negative when q is small. In the scaled co-ordinates the Takens–Bogdanov point is located at $(\hat{\mu}_{TB}, \hat{\beta}_{TB}) = (2(2 - \sigma_{\max})/\sigma_{\max}, -2\sigma_{\max})$ which, when $\sigma_{\max} = 1$, is exactly the same as for the ALL model. Similarly, on setting $\sigma_{\max} = 1$, the curves along which the pitchfork and Hopf bifurcations lie reduce to those for the ALL model. The pitchfork bifurcation occurs at

$$\hat{\beta} = \frac{\sigma_{\max}^2}{\sigma_{\max} - 2}\hat{\mu},$$

and the Hopf bifurcation lies in exactly the same place as for the product cross-immunity case:

$$\hat{\beta} = \frac{4(\sigma_{\max} - 2)}{\hat{\mu}}, \qquad \text{when} \qquad \hat{\mu} > \hat{\mu}_{TB}.$$

On computing the coefficients P and Q we find that $Q = -\sigma_{\max}/32$ as we would expect, but $P = 0$; this model shares the degeneracy in the computation of P that was noted in section 3.2 for the ALL model. We have not computed the higher-order contributions to P in this case, but confidently expect to find that $P > 0$ at higher orders, as was the case for the ALL model.

In summary, the source of the oscillations in the model described by Gomes et al. [10] is exactly the same as in the ALL model. From the bifurcation theory viewpoint, the point $p = 0, \sigma_{\max} = 1$ has degenerate features, and the behaviour at fixed $\sigma_{\max}$ as p is varied is qualitatively as was found for the ALL model with varying σ. For different choices of $\sigma_{\max}$ the bifurcation curves in the (r, p) plane shift a little, but the structure is unchanged; for r near 1 and $p > 0$ the fully-symmetric equilibrium is stable while for $p < 0$ the edge equilibrium is stable. For values of p near zero we have proved the existence of oscillations, and the bifurcation structure here is exactly that of figure 6, with appropriate re-labelling of the y-axis.

4. Discussion and conclusions

In this paper we have developed a mathematical approach to understanding the complex dynamics inherent in models of multiple disease strains. This approach exploited the small parameter $e = b/(b + \nu)$ which can be interpreted as the ratio of the average duration of an infection to a host individual's lifetime. Asymptotic analysis in the limit of small e was first applied to a generalised model of the dynamics of four competing disease strains, and enabled us to describe the possibility of oscillatory dynamics using well-known results describing Takens–Bogdanov bifurcations.

Four specific models were then analysed within this framework. Each of these was a special case of the generalised model, and different specific models contained different motivating assumptions. The bifurcation-theoretic results for the generalised model explain, in a consistent and unifying way, the presence of oscillations in all but one of these specific models, and provide a mathematical reason for the absence of oscillations in the GS model.

It remains to fully explain which biological assumptions lead to models that may sustain oscillations. This does not immediately follow on from the mathematical analysis as the expressions generated do not easily fit with any biologically-interpretable measures. One might suspect that perhaps all status-based models do not exhibit oscillations, most likely through the assumption of polarised immunity used in the GS model (any given host is either totally susceptible or totally immune). However, models corresponding to sets of coefficients which are very close to those that specify the GS model do show oscillations. For example reducing τ_1 to less than unity, keeping all other coefficients as specified in table 1, is enough to excite oscillations in a small region around $\sigma = 1/2$, see figure 11. Although this system is no longer one of polarised immunity, it is only a small deviation from the GS model. If we cannot interpret why this small alteration changes the dynamics, we do not have a complete understanding of what leads to the oscillations.

However, oscillations in four-strain models are not what is driving strain replacement observable on the time-scale of a host lifetime. When oscillations are present, section 2.4.2 shows that, except extremely close to the Takens-Bogdanov point, the period of oscillation is roughly of the same order as host lifetime. This is fairly straightforward to understand: once the susceptibles are depleted for one strain, the only way for them to recruit new hosts is through new births. The frequency of the oscillations is limited by the population birthrate.

That the strain variation observed in these models operates on a scale longer than an individual's lifetime can be seen from figure 8. Although trajectories take only a short time to move most of the distance around the orbit, they slow down close to the edge equilibria which adds an order-one contribution to the total period. Hence the flip between different pairs of strains takes many host lifetimes (many centuries in human terms). So we conclude that the cycles possible in four strain models are qualitatively different to any observed cycling in strains that occurs more rapidly than a typical host lifetime, for example switches between subtypes of influenza which occur every year or two.

For cycles in models containing more than four strains, we still expect that the rise and fall in prevalence for each strain does not occur on a timescale much faster than the lifetime of an individual host. However, through the combination of many strains rising and falling it is possible to observe a rapid succession of different prevalent strains. For example, in figure 2B of [13] where there are 8 strains, the host lifetime is about 50 years, and we see the most prevalent strains changing roughly every 5 years. The maintenance of fast dynamics is reliant on the availability of many strains.

Irregular, and possibly chaotic, oscillations have been observed by several authors, but always at larger values of r than those that we consider in this paper. It seems highly likely that these are due to secondary bifurcations, away from $r = 1$,

of the periodic oscillations analysed in this paper. The lack of chaotic oscillations observed close to $r = 1$ in previous numerical simulations is a further point of agreement between this analysis and previous papers.

Another possible mathematical approach to the structure of these models of multiple strains is that of a system containing fast–slow dynamics (dynamics on two timescales). When e is small, the force of infection variables Λ_j evolve on a fast timescale ($\dot{\Lambda}_j \sim 1/e$), whereas the proportions of individuals in each class S_J evolve on a much slower (order 1) timescale. This structure occurs in a wide range of mathematical models of biological systems, and could possibly be exploited here.

Several directions for future work present themselves; we briefly discuss three of these here. Firstly, we would like to explore how similar analysis can be used to understand the complex dynamics observed in models containing more than four strains, see for example [10]. The arrangement of the strains in a 'ring' must affect the dynamics of the system; general properties of rings of, say, coupled oscillators are well-known, and may be of use in these problems. In particular, the symmetries of the set of strains arranged in a ring can be described as a wreath product [4]; a wreath product is the symmetry group that naturally results from combining an 'local' symmetry for each strain (which is due to the fact that if any strain is not present initially, then it cannot be created by the dynamics) with the 'global' symmetry of the arrangement of the strains. In the case of n strains spaced equally around a circle, this global group is the symmetries of an n-gon, usually denoted D_n.

Secondly, it is of interest to investigate how robust the dynamics of these models are when the global symmetries, which arise from treating all strains equally, are broken, for example by varying the reproduction ratio r from one strain to another. We note that this point was investigated briefly by Andreasen et al [1] for their model. It is possible, although we believe not likely, that such symmetry-breaking effects may lead to chaotic dynamics for r close to unity. Work on this issue is in progress.

Thirdly, as mentioned above, it is important to give a biological explanation of the assumptions that lead to oscillations. We have given a comprehensive mathematical description of the generation of oscillations and described previous models in a unified framework. Our suspicion is that the polarised immunity is a key factor in distinguishing different models. The mathematical analysis presented here is likely to be of use in any further work in this direction.

Ultimately, as immunological and genetic advances are made, there is increasing demand for mathematical models that are able to describe and eventually to predict these highly complex systems of multiple strains. To apply models to systems with a large number of strains, and to systems with complex strain structure, the modeller will be forced to choose a convenient set of assumptions. It is imperative that the impact of these assumptions is fully understood. We specifically chose systems of only four strains for analysis since even here there are differences in existing multiple strain models. As we look to more intricate strain models, understanding gained from explorations of simpler systems will be invaluable. This may give us confidence in some model choices and enable an understanding of the potential hazards involved in other selections. To this end, further study of systems of limited numbers of strains should be pursued.

Appendix

In this appendix we sketch some of the details of the reduction procedure through which we derive the simplified equations describing the Takens–Bogdanov bifurcation (53) - (54), starting from (40) - (47).

Step 1. We rescale the variables Λ_S, Λ_D, ϕ_S, ϕ_D, U, S_S, S_D and S_{12}:

$$\begin{aligned} \Lambda_S &= \sqrt{e}\hat{\Lambda}_S, & \phi_S &= \sqrt{e}\hat{\phi}_S, & S_S &= \sqrt{e}\hat{S}_S, \\ \Lambda_D &= \sqrt{e}\hat{\Lambda}_D, & \phi_D &= e\hat{\phi}_D, & S_D &= \sqrt{e}\hat{S}_D, \\ U &= \sqrt{e}\hat{U}, & S_{12} &= \sqrt{e}\hat{S}_{12}, \end{aligned}$$

and also write $r = 1+\hat{\mu}\sqrt{e}$. From section 2.4.2 we also require $p_1 - p_2 \sim \sqrt{e}$ for the Takens–Bogdanov point to exist within this scaling, hence we write $p_2 - p_1 = \hat{\beta}\sqrt{e}$. We now drop these hats and refer below exclusively to the scaled versions of the variables.

Step 2. We make use of the fact that the Takens–Bogdanov bifurcation occurs in the sub-system involving only the Λ_D, ϕ_D and S_D variables. We perform a centre-manifold reduction to write the variables $\{\Lambda_S, \phi_S, U, S_S, S_{12}\}$ in terms of $\{\Lambda_D, \phi_D, S_D\}$ and hence eliminate them from the evolution equations for $\{\Lambda_D, \phi_D, S_D\}$. In performing this reduction we keep the linear terms and only the lowest-order nonlinearities. We first use the $\dot{U}$, $\dot{S}_S$ and $\dot{S}_{12}$ equations (44), (45) and (47) to write U, S_S and S_{12} in terms of Λ_S:

$$U = 2\Lambda_S[1 - 2\sqrt{e}\Lambda_S + 4e\Lambda_S^2] + \mathcal{O}(e\sqrt{e}), \tag{59}$$

$$S_S = a_1\Lambda_S + \sqrt{e}[-2a_1\Lambda_S^2 + (b_0 - c_0/2)\Lambda_D S_D - a_1(b_0 + c_0/2)\Lambda_S^2] + \mathcal{O}(e), \tag{60}$$

$$S_{12} = a_2\Lambda_S + \sqrt{e}[-a_2 d_0\Lambda_S^2 - 2a_2\Lambda_S^2 + a_1 b_1\Lambda_S^2/2 - b_1\Lambda_D S_D/2] + \mathcal{O}(e). \tag{61}$$

Now we turn our attention to the $\dot{\Lambda}_S$ and $\dot{\phi}_S$ equations (40) and (42). We expand Λ_S and ϕ_S in powers of $\sqrt{e}$:

$$\Lambda_S = -\frac{\mu}{p_1} + x_1\sqrt{e} + x_2 e + \mathcal{O}(e\sqrt{e}), \tag{62}$$

$$\phi_S = 2\mu + \sqrt{e}\tilde{\phi}_1 + \mathcal{O}(e), \tag{63}$$

and substitute these into (40) to obtain

$$\tilde{\phi}_1 = \frac{p_1}{\mu}\Lambda_D\phi_D - 2\mu^2. \tag{64}$$

We can now substitute the expansions (62) and (63) along with the result (64) into the $\dot{\phi}_S$ equation (42) to find

$$x_1 = -\frac{1}{2\mu}\Lambda_D\phi_D - \frac{p_4}{2p_1}\Lambda_D S_D + \text{linear terms}. \tag{65}$$

The linear terms in x_1 may be ignored since they contribute only linear terms of size $\sqrt{e}$ to Λ_S, and we are interested in finding the lowest-order nonlinear terms. We can now use (64) and (65) to give the leading-order nonlinearities in (62) and (63):

$$\Lambda_S = -\frac{\mu}{p_1} - \sqrt{e}\left[\frac{1}{2\mu}\Lambda_D S_D + \frac{p_4}{2p_1}\Lambda_D S_D\right] + \mathcal{O}(e),$$
$$\phi_S = 2\mu\sqrt{e} + e\left[\frac{p_1}{\mu}\Lambda_D\phi_D - 2\mu^2\right] + \mathcal{O}(e\sqrt{e}),$$

and hence find the leading-order nonlinearities in the expressions for U, S_S and S_{12}:

$$U = -\frac{2\mu}{p_1} + \sqrt{e}\left[2x_1 - \frac{4\mu^2}{p_1^2}\right] + \mathcal{O}(e),$$
$$S_S = -\frac{a_1\mu}{p_1} + \sqrt{e}\left[a_1x_1 - \frac{2a_1\mu^2}{p_1^2} + (b_0 + c_0/2)\Lambda_D S_D - \frac{a_1\mu^2(b_0 + c_0/2)}{p_1^2}\right] + \mathcal{O}(e),$$
$$S_{12} = -\frac{a_2\mu}{p_1} + \sqrt{e}\left[a_2x_1 - \frac{a_2d_0\mu^2}{p_1^2} - \frac{2a_2\mu^2}{p_1^2} + \frac{a_1b_1\mu^2}{2p_1^2} - \frac{b_1\Lambda_D S_D}{2}\right] + \mathcal{O}(e).$$

The last part of this step is to substitute these expressions into the $\dot{\Lambda}_D$, $\dot{\phi}_D$ and $\dot{S}_D$ equations. This yields three ODEs involving only Λ_D, ϕ_D and S_D:

$$\begin{pmatrix}\dot{\Lambda}_D\\ \dot{\phi}_D\\ \dot{S}_D\end{pmatrix} = \begin{pmatrix}0 & \frac{\mu}{2p_1} & 0\\ \hat{c} & -1 & \frac{p_5\mu}{p_1}\\ a_1 & 0 & -1\end{pmatrix}\begin{pmatrix}\Lambda_D\\ \phi_D\\ S_D\end{pmatrix} + \begin{pmatrix}-\frac{p_1}{2\mu}\Lambda_D^2\phi_D + \mathcal{O}(\sqrt{e})\\ \sqrt{e}(A\Lambda_D^2\phi_D + B\Lambda_D^2 S_D + C\Lambda_D S_D^2 + D\Lambda_D\phi_D S_D) + \mathcal{O}(e)\\ \mathcal{O}(e)\end{pmatrix} \quad (66)$$

where

$$\hat{c} = \beta + \frac{\mu}{p_1}[a_1p_6 + 2a_2(\tau_3f_0 - \tau_5d_1) - 2f_0 - f_0p_1],$$
$$A = \frac{-f_0(2 + p_1) + a_1p_6}{2\mu},$$
$$B = -\frac{f_0p_4}{p_1} + \frac{a_1p_4p_6}{2p_1} - p_6(b_0 - c_0/2) + (\tau_3f_0 - \tau_5d_1)\left(\frac{a_2p_4}{p_1} + b_1\right),$$
$$C = \frac{p_4p_5}{2p_1},$$
$$D = \frac{p_5}{2\mu}.$$

Note that the linear terms (taking the scalings of Λ_D, ϕ_D and S_D into account) agree with the linear analysis of section 2.4.2. Here we would usually want only to

keep the leading-order nonlinearities, but it turns out that, to avoid degeneracies in the final stage of the reduction, it is necessary to keep the $\mathcal{O}(\sqrt{e})$ nonlinearities in the $\dot{\phi}_D$ equation as well as the nonlinear term in the $\dot{\Lambda}_D$ equation.

Step 3. The final step is to perform a linear change of variables to reduce the third-order set of equations in $\{\Lambda_D, \phi_D, S_D\}$ to a second-order set which are in a standard, simplest possible, form. The resulting coefficient values of the nonlinear terms in this second-order system select which particular case of the Takens–Bogdanov bifurcation applies, and from the associated theory we can construct the relevant bifurcation diagrams. This analysis is summarised in section 2.5. The linear change of variables is given by

$$\begin{aligned} x &= -\Lambda_D - S_D, \\ y &= -\Lambda_D + \phi_D + S_D, \\ z &= -\Lambda_D - \phi_D + S_D. \end{aligned}$$

At the Takens–Bogdanov point $(\mu, \beta) = (\mu_{TB}, \beta_{TB})$, where the linear part of (66) has two zero eigenvalues and one negative one, x and y are slowly-evolving variables on the centre manifold and z decays fast and so may be eliminated. This leaves two ODEs for x and y which contain a large number of cubic nonlinearities:

$$\begin{aligned} \dot{x} &= y - N_1, \\ \dot{y} &= N_2 - N_1, \end{aligned}$$

where N_1 and N_2 are the nonlinear terms in the $\dot{\Lambda}_D$ and $\dot{\phi}_D$ components of (66), respectively. These equations for $\dot{x}$ and $\dot{y}$ can be further simplified by making a near-identity change of co-ordinates $(x', y') = (x, y) +$ cubic terms which can be chosen to eliminate all but two of the cubic terms in the resulting equations for $\dot{x}'$ and $\dot{y}'$. Dropping the primes, the ODEs for x and y are now in the form

$$\dot{x} = y, \tag{67}$$

$$\dot{y} = Px^3 + Qx^2 y. \tag{68}$$

Note that, since

$$N_1 \sim \Lambda_D^2 \phi_D \equiv \left(\frac{x}{2} + \frac{y}{4}\right)^2 \frac{y}{2},$$

omitting N_2 would lead to equations for $\dot{x}$ and $\dot{y}$ which contain no x^3 term. The equations (67) - (68) apply *exactly at* the bifurcation point. To investigate the dynamics in a neighbourhood of the Takens–Bogdanov point we include two linear unfolding terms in the standard manner, and analyse the resulting system

$$\dot{x} = y, \tag{69}$$

$$\dot{y} = -\lambda x + \kappa y + Px^3 + Qx^2 y, \tag{70}$$

which is exactly (53) - (54). The analysis of these ODEs is summarised in section 2.5; a full description can also be found in [12] following earlier work by [20] and [16].

The lack of quadratic terms is not a coincidence; it is due to the symmetry of the original generalised model under permutation of the indices $\{1, 2, 3, 4\} \to \{2, 3, 4, 1\}$. This symmetry implies that, in the 'sum and difference' co-ordinates (recall the restriction to the 'diagonal' subspace where $\Lambda_1 = \Lambda_3$, $\Lambda_2 = \Lambda_4$ etc), the equations are required to be invariant under the transformation

$$\{\Lambda_S, \Lambda_D, \phi_S, \phi_D, U, S_S, S_D, S_{12}\} \to \{\Lambda_S, -\Lambda_D, \phi_S, -\phi_D, U, S_S, -S_D, S_{12}\}.$$

Furthermore, x, y and z are linear combinations of Λ_D, ϕ_D and S_D, so the resulting reduced equations can contain only terms of odd total order, and they are invariant under the symmetry $(x, y) \to (-x, -y)$. This symmetric version of the Takens–Bogdanov bifurcation (often referred to as the $\mathbb{Z}_2$-symmetric Takens–Bogdanov bifurcation) is well-known since it occurs naturally in analyses of various fluid mechanical problems, see for example [18].

Acknowledgements. We have enjoyed useful discussions about this work with Bryan Grenfell, Gabriela Gomes and Michael Proctor. We are also very grateful for many constructive comments from the referees. JHPD is very grateful for the hospitality of the Department of Mathematics and Statistics at the University of Otago where part of this work was carried out. Both authors are funded by Trinity College, Cambridge.

References

1. Andreasen, V., Lin, J., Levin, S.A.: The dynamics of cocirculating influenza strains conferring partial cross-immunity. *J. Math. Biol.* **35**, 825–842 (1997)
2. Castillo–Chavez, C., Hethcote, H.W., Andreasen, V., Levin, S.A., Liu, W.M.: Epidemiological models with age structure, proportionate mixing and cross-immunity. J. Math. Biol. **27**, 233–258 (1989)
3. Cox, N.J., Subbarao, K.: Global Epidemiology of Influenza: Past and Present. Annu. Rev. Med. **51**, 407–421
4. Dionne, B., Golubitsky, M., Stewart, I.: Coupled cells with internal symmetry: I. Wreath products. *Nonlinearity* **9**, 559–574 (1996)
5. Doedel, E., Champneys, A., Fairgrieve, T., Kuznetsov, Y., Sandstede, B., Wang, X.: *AUTO97: Continuation and bifurcation software for ordinary differential equations.* Available via FTP from directory `pub/doedel/auto` at `ftp.cs.concordia.ca` (1997)
6. Ferguson, N., Andreasen, V.: The influence of different forms of cross-protective immunity on the population dynamics of antigenically diverse pathogens. In: Blower, S., Castillo-Chavez, C., Cooke, K.L., Kirschner, D., van der Driessche, P. (eds) Mathematical Approaches for Emerging and Re-emerging Infections, IMA Volumes in Mathematics and its Applications. Springer, New York. In press. 2001
7. Ferguson, N.M., Anderson, R.M., Gupta, S.: The effect of antibody-dependent enhancement on the transmission dynamics and persistence of multiple-strain pathogens. Proc. Natl. Acad. Sci. USA **96**(2), 790–794 (1999)
8. Gog, J.R., Swinton, J.: A status-based approach to multiple strain dynamics. J. Math. Biol. **44**, 169–18 (2002)
9. Golubitsky, M., Stewart, I.: *The Symmetry Perspective*. Progress in Mathematics Volume 200. Birkhäuser, 2001
10. Gomes, M.G.M., Medley, G., Nokes, D.J.: On the determinants of population structure in antigenically diverse pathogens. Proc. R. Soc. Lond. B **269**, 227–233 (2002)

11. Grenfell, B.T., Gog, J.R.: Pathogen strains: no joke. Trends Ecol. Evol. **16**(6), 2001
12. Guckenheimer, J., Holmes, P.: *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields.* Springer: New York, 1983
13. Gupta, S., Ferguson, N., Anderson, R.: Chaos, persistence and evolution of strain structure in antigenically diverse infectious agents. *Science* **280**, 912–915 (1998)
14. Gupta, S., Trenholme, K., Anderson, R.M., Day, K.P.: Antigenic diversity and the transmission dynamics of *Plasmodium falciparum*. Science **263**, 961–963 (1994)
15. Gupta, S., Anderson, R.M.: Population structure of pathogens: The role of immune selection. Parasitol. Today **15**(12), 497–501 (1999)
16. Holmes, P.J., Rand, D.A.: Phase portraits and bifurcations of the nonlinear oscillator $\ddot{x} + (\alpha + \gamma x^2)\dot{x} + \beta x + \delta x^3 = 0$. *Int. J. Nonlinear Mech.* **15**, 449–458 (1980)
17. Knobloch, E.: On the degenerate Hopf bifurcation with $O(2)$ symmetry. In: Golubitsky, M., Guckenheimer, J. (eds) Multiparameter Bifurcation Theory, volume **56** of *Contemporary Mathematics* pp. 193–201. American Mathematical Society, Providence, R.I., USA, 1986
18. Knobloch, E., Proctor, M.R.E.: Nonlinear periodic convection in double-diffusive systems. *J. Fluid Mech.* **108**, 291–316 (1981)
19. Swift, J.W.: Bifurcation and Symmetry in Convection. PhD thesis, University of California, Berkeley, 1984
20. Takens, F.: Forced oscillations and bifurcations. *Comm. Math. Inst., Rijksuniversiteit Utrecht* **3**, 1–59 (1974)