JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2010) **7**, 873–885
doi:10.1098/rsif.2009.0386
Published online 5 November 2009

The construction of next-generation matrices for compartmental epidemic models

O. Diekmann[1], J. A. P. Heesterbeek[2,*], and M. G. Roberts[3]

[1]*Department of Mathematics, Utrecht University, Budapestlaan 6, 3584 CD, Utrecht, The Netherlands*
[2]*Faculty of Veterinary Medicine, Utrecht University, Yalelaan 7, 3584 CL, Utrecht, The Netherlands*
[3]*Centre for Mathematical Biology, Institute of Information & Mathematical Sciences, Massey University, Private Bag 102 904, North Shore Mail Centre, Auckland, New Zealand*

The basic reproduction number $\mathcal{R}_0$ is arguably the most important quantity in infectious disease epidemiology. The next-generation matrix (NGM) is the natural basis for the definition and calculation of $\mathcal{R}_0$ where finitely many different categories of individuals are recognized. We clear up confusion that has been around in the literature concerning the construction of this matrix, specifically for the most frequently used so-called compartmental models. We present a detailed easy recipe for the construction of the NGM from basic ingredients derived directly from the specifications of the model. We show that two related matrices exist which we define to be the *NGM with large domain* and the *NGM with small domain*. The three matrices together reflect the range of possibilities encountered in the literature for the characterization of $\mathcal{R}_0$. We show how they are connected and how their construction follows from the basic model ingredients, and establish that they have the same non-zero eigenvalues, the largest of which is the basic reproduction number $\mathcal{R}_0$. Although we present formal recipes based on linear algebra, we encourage the construction of the NGM by way of direct epidemiological reasoning, using the clear interpretation of the elements of the NGM and of the model ingredients. We present a selection of examples as a practical guide to our methods. In the appendix we present an elementary but complete proof that $\mathcal{R}_0$ defined as the dominant eigenvalue of the NGM for compartmental systems and the Malthusian parameter r, the real-time exponential growth rate in the early phase of an outbreak, are connected by the properties that $\mathcal{R}_0 > 1$ if and only if $r > 0$, and $\mathcal{R}_0 = 1$ if and only if $r = 0$.

Keywords: basic reproduction number; next-generation matrix; epidemiological model

1. INTRODUCTION

The basic reproduction number $\mathcal{R}_0$ is arguably the most important quantity in infectious disease epidemiology. It is among the quantities most urgently estimated for emerging infectious diseases in outbreak situations, and its value provides insight when designing control interventions for established infections. From a theoretical point of view $\mathcal{R}_0$ plays a vital role in the analysis of, and consequent insight from, infectious disease models. There is hardly a paper on dynamic epidemiological models in the literature where $\mathcal{R}_0$ does not play a role. $\mathcal{R}_0$ is defined as the average number of new cases of an infection caused by one typical infected individual, in a population consisting of susceptibles only.[1]

It has been shown that $\mathcal{R}_0$ is *mathematically* charactrized by regarding infection transmission as a 'demographic process', where producing offspring is not seen as giving birth in the demographic sense, but as causing a new infection through transmission (we will refer to this as an '*epidemiological birth*'). In a natural way this leads to viewing the infection process in terms of consecutive 'generations of infected individuals', in complete analogy to demographic generations. Subsequent generations growing in size then indicate a growing population (i.e. an epidemic), and the growth factor per generation indicates the potential for growth. In a natural way this growth factor is then the mathematical characterization of $\mathcal{R}_0$ (Diekmann *et al.* 1990).

As a rule, several traits of individuals are epidemiologically relevant in an infectious agent/host system: for example age, sex, species. We will only regard the case where these traits divide the population into a finite number of discrete categories. One can then define

*Author for correspondence (j.a.p.heesterbeek@uu.nl).
All authors contributed equally to the study, the order is alphabetical.

Received 1 September 2009
Accepted 16 October 2009

a matrix that relates the numbers of newly infected individuals in the various categories in consecutive generations. This matrix, usually denoted by $\mathbf{K}$, is called the *next-generation matrix* (NGM); it was introduced in Diekmann *et al.* (1990) who proposed to define $\mathcal{R}_0$ as the dominant eigenvalue of $\mathbf{K}$. In this paper, we show how to construct the NGM for any such system. We relate the structure of the NGM to its epidemiological interpretation, and use this interpretation to extract relevant information from the matrix in a systematic manner.

Compartmental models are the most frequently used type of epidemic model. In this class of models, individuals can be in a finite number of discrete states. Some of these states are simply labels that specify the various traits of individuals. Of these, some will be changing with time, such as age class, and others will be fixed, such as sex or species. Other states indicate the progress of an infection: for example, an individual can upon becoming infected, typically first enter a state of latency, then progress to a state of infectiousness, and then lose infected status to progress to a recovered/immune state. With each state one can associate the subpopulation of individuals who are in that particular state at the given time (e.g. a female in a latent state of infection). Often the same symbol is used as a label for a state and to denote the corresponding subpopulation size, either as a fraction or as a number (e.g. I or Y for individuals in an infectious state). The dynamics are generated by a system of nonlinear ordinary differential equations (ODEs) that describes the change with time for all subpopulation sizes. For the computation of $\mathcal{R}_0$ we only regard the states that apply to infected individuals.

To calculate $\mathcal{R}_0$ one begins with those equations of the ODE system that describe the production of new infections and changes in state among infected individuals. We will refer to the set of such equations as the *infected subsystem*. The first step is to linearize the infected subsystem of nonlinear ODEs about the infection-free steady state that, as a rule, exists. Epidemiologically the linearization reflects that $\mathcal{R}_0$ characterizes the potential for initial spread of an infectious agent when it is introduced into a fully susceptible population, and that we assume that the change in the susceptible population is negligible during the initial spread. This linearized infected subsystem is the starting point of our calculations.

Any linear system of ODEs is described by a matrix, usually called the Jacobian matrix when derived by linearization of the original nonlinear ODE system. Our aim is to relate the structure of this matrix to the epidemiological interpretation. In particular, we explain how one can determine $\mathcal{R}_0$ by first decomposing the matrix as $\mathbf{T}+\mathbf{\Sigma}$, where $\mathbf{T}$ is the trans*mission* part, describing the production of new infections, and $\mathbf{\Sigma}$ is the trans*ition* part, describing changes in state (including removal by death or the acquisition of immunity). We next compute the dominant eigenvalue, or more precisely the spectral radius ρ, of the matrix $-\mathbf{T}\mathbf{\Sigma}^{-1}$ (note the minus sign in front of $\mathbf{T}$). This decomposition into $\mathbf{T}$ and $\mathbf{\Sigma}$ was first described in Diekmann & Heesterbeek (2000, pp. 105–107) and later in Van den Driessche & Watmough (2002), but does not typically lead to the NGM as introduced in Diekmann *et al.* (1990; and elaborated in Diekmann & Heesterbeek (2000, ch. 5)), which is the basis for the definition of $\mathcal{R}_0$. This is because the decomposition relates to the expected offspring of individuals of any state and not just epidemiological newborns (i.e. new infections). For example, a latency state and a consecutive infectious state are both infected states, but the change from latency to infectiousness does not involve a new infection occurring, but rather an already established infection moving to a different infection stage. This has led to confusion as others have tried to reconcile the appealing linear algebra approach with the original NGM $\mathbf{K}$ and its interpretation. To make the distinction clear and remove confusion, we will call the matrix $\mathbf{K}_{\mathrm{L}} := -\mathbf{T}\mathbf{\Sigma}^{-1}$ the *NGM with Large domain* (hence the subscript 'L'). We will show that $\rho(\mathbf{K}_{\mathrm{L}}) = \rho(\mathbf{K})$.

We will show how one can easily find the NGM $\mathbf{K}$ from the NGM with large domain $\mathbf{K}_{\mathrm{L}}$. This is important because very often (indeed almost always) $\mathbf{K}$ has a dimension which is lower than that of $\mathbf{K}_{\mathrm{L}}$, making the computation of $\mathcal{R}_0$ from $\mathbf{K}$ easier and increasing the possibility of obtaining an explicit expression. The reason for this is that there are usually but a few states that can be entered through epidemiological birth among the total number of states in the system. The NGM with large domain typically uses the dynamics of (many) more states than the NGM to describe the evolution of infection generations. Because the epidemiological births represent states that individuals can have immediately following their infection, we will call these *states-at-infection*.[2] Only the states-at-infection are involved in the action of $\mathbf{K}$, and hence in the computation of $\mathcal{R}_0$. By regarding the matrix $\mathbf{T}$ we show how one can easily determine the states-at-infection; a simple matrix calculation using $\mathbf{T}$ and $\mathbf{\Sigma}$ then leads to $\mathbf{K}$.

In some situations a further reduction in dimension is possible. This is the case when $\det \mathbf{K} = 0$. Typically this is when the incidences corresponding to two or more different states-at-infection occur in a fixed (i.e. time independent) ratio. We call the lower-dimensional matrix the *NGM with Small domain*, and denote it by $\mathbf{K}_{\mathrm{S}}$. We will show how to compute the smaller matrix from the basic ingredients in $\mathbf{T}$ and $\mathbf{\Sigma}$, and that the spectral radius of $\mathbf{K}_{\mathrm{S}}$ is equal to that of $\mathbf{K}$.

Experienced modellers can often jump directly from the model specification to the NGM, without going through the formalities of the linear algebra involved. Even though the construction of $\mathbf{K}_{\mathrm{L}}$ is an easy exercise from the linear algebra perspective, and $\mathbf{K}$ may be derived from $\mathbf{K}_{\mathrm{L}}$ via a linear transformation, the construction of $\mathbf{K}$ is even easier if one is guided by the epidemiological interpretation. This is possible because of the clear biological meaning of the elements of $\mathbf{K}$. The element K_{ij} is the expected number of new cases with state-at-infection i, generated by one individual who has just been born (epidemiologically speaking) in state-at-infection j. Throughout this paper we will emphasize this intuitive approach for all examples used, in the hope that less experienced modellers are able to gain insight into deriving the NGM in this systematic and rigorous, yet biological, manner.

The reduction process sketched above often leads to an explicit formula for $\mathcal{R}_0$ or, at least, to an eigenvalue problem with lowest possible dimension (given the specified biology of the system). This is one of the reasons why researchers compute $\mathcal{R}_0$ and not the intrinsic rate of natural increase (Malthusian parameter) r, which would otherwise serve equally well to characterize the potential for initial spread. In general, there is no explicit relation between the value of $\mathcal{R}_0$ and the value of r, in the sense that, for example, infections with a high $\mathcal{R}_0$ do not automatically lead to fast exponential increase of incidence.

However, the magnitude of $\mathcal{R}_0$ does reveal the sign of r because the following holds: $\mathcal{R}_0 > 1$ if and only if $r > 0$, and $\mathcal{R}_0 = 1$ if and only if $r = 0$ (and hence one also has $\mathcal{R}_0 < 1$ if and only if $r < 0$). In the appendix we provide an elementary but detailed proof of this correspondence. The proof originally given in Diekmann & Heesterbeek (2000) is incomplete, as pointed out in Thieme (2009; H. R. Thieme 2009, personal communication). It is this sign equivalence that validates the use of the generation-based approach to characterize $\mathcal{R}_0$ and hence the theory of the NGM. This relation with r establishes that $\mathcal{R}_0 > 1$ implies instability of the infection-free steady state of the ODE system, and $\mathcal{R}_0 < 1$ implies stability. This is helpful because, in a model setting, it is often possible to derive a formula for $\mathcal{R}_0$, whereas r is only implicitly defined.

2. MOTIVATING EXAMPLES

To illustrate the various NGMs that were introduced in §1, we construct two connected examples for pedagogical purposes only. In §3 we present formal recipes to derive the various NGMs in general. In the present motivational section we explain the foundations for the steps in the recipes in the context of the examples. Both examples relate to a compartmental *SEI* model where there are two categories of individuals in the population. For the first example the only epidemiological difference between the categories is the time that the individuals spend in the latent phase following exposure to infection. The second example is an extension of this model where the two categories respond differently to infection throughout their life (susceptibility, latency, infectivity).

2.1. An SEI model with two latent categories

Consider a system with the following states: S susceptible; E_1 latently infected of category 1; E_2 latently infected of category 2; I infectious; and R recovered/removed/immune. As usual, the letters for the states also indicate the size of the subpopulation in that state, where 'size' in our case is the number of individuals in that state. The idea behind this system might be that categories 1 and 2 represent individuals who, once infected, progress to infectiousness at different rates. For this model, we assume that the trait that causes this difference in disease progression does not manifest itself as a difference in susceptibility, so there is only one S state. We assume that there is a fixed ratio of the two categories in the population, $p : 1 - p$, hence susceptibles enter the E_1 and E_2 states in that fixed ratio following exposure to infection. Let β be the transmission rate, μ the *per capita* birth and death rates, ν_1 and ν_2 the rates of leaving the respective latency states, and γ the rate of leaving the infectious state. The equations are

$$\dot{S} = \mu N - \beta \frac{SI}{N} - \mu S, \tag{2.1}$$

$$\dot{E}_1 = p\beta \frac{SI}{N} - (\nu_1 + \mu)E_1, \tag{2.2}$$

$$\dot{E}_2 = (1-p)\beta \frac{SI}{N} - (\nu_2 + \mu)E_2, \tag{2.3}$$

$$\dot{I} = \nu_1 E_1 + \nu_2 E_2 - (\gamma + \mu)I \tag{2.4}$$

and

$$\dot{R} = \gamma I - \mu R, \tag{2.5}$$

with $N = S + E_1 + E_2 + I + R$. This system has three infected states, E_1, E_2, and I; and two uninfected states, S and R. Although there are five states in the model, it is four-dimensional as the total population size is constant. At the infection-free steady state $E_1 = E_2 = I = R = 0$, hence $S = N$. The only occurrence of the variable S in equations (2.2)–(2.5), either directly or implicitly via N, is through the term $\beta SI/N$ in equations (2.2) and (2.3) which becomes βI when we set $S = N$. Hence the linearization of equations (2.2)–(2.4) is closed, in that it does not involve the deviation of S from its steady-state value. Also, R does not appear in equations (2.2)–(2.4), and for small (E_1, E_2, I) we have the linear system

$$\dot{E}_1 = p\beta I - (\nu_1 + \mu)E_1, \tag{2.6}$$

$$\dot{E}_2 = (1-p)\beta I - (\nu_2 + \mu)E_2 \tag{2.7}$$

and

$$\dot{I} = \nu_1 E_1 + \nu_2 E_2 - (\gamma + \mu)I. \tag{2.8}$$

We will refer to the ODEs (2.6)–(2.8) as the linearized *infection subsystem*, as it only describes the production of new infecteds and changes in the states of already existing infecteds.

If we set $\mathbf{x} = (E_1, E_2, I)'$, where the prime denotes transpose, we now want to write the linearized infection subsystem in the form

$$\dot{\mathbf{x}} = (\mathbf{T} + \mathbf{\Sigma})\mathbf{x}. \tag{2.9}$$

The matrix $\mathbf{T}$ corresponds to *transmissions* and the matrix $\mathbf{\Sigma}$ to *transitions*. In this paper, we include death in the transition matrix to keep the notation simple (contrast with Diekmann & Heesterbeek 2000). Hence, all epidemiological events that lead to new infections are incorporated in the model via $\mathbf{T}$, and all other events via $\mathbf{\Sigma}$. Progress to either death or immunity guarantees that $\mathbf{\Sigma}$ is invertible.

Our example, described by the subsystem (2.6)–(2.8), is three-dimensional and hence the transmission and transition matrices in the corresponding description (2.9) are also three-dimensional. They are obtained from system (2.6)–(2.8) by separating the transmission events from other events. If we refer to the infected states with indices i and j, with $i,j \in 1, 2, 3$, then the entry T_{ij} is the rate at which individuals

in infected state j give rise to individuals in infected state i, in the linearized system. So $T_{ij}=0$ when no new cases produced by an individual in infected state j can be in infected state i immediately after infection. Regarding the linearized subsystem (2.6)–(2.8) we obtain

$$\mathbf{T}=\begin{pmatrix}0 & 0 & p\beta\\ 0 & 0 & (1-p)\beta\\ 0 & 0 & 0\end{pmatrix}$$

and

$$\Sigma=\begin{pmatrix}-(\nu_1+\mu) & 0 & 0\\ 0 & -(\nu_2+\mu) & 0\\ \nu_1 & \nu_2 & -(\gamma+\mu)\end{pmatrix}.$$

Hence the NGM with large domain $\mathbf{K}_{\mathrm{L}}$ is three-dimensional and given by (note the essential minus sign)

$$\begin{aligned}\mathbf{K}_{\mathrm{L}}=-\mathbf{T}\Sigma^{-1}&=\begin{pmatrix}0 & 0 & p\beta\\ 0 & 0 & (1-p)\beta\\ 0 & 0 & 0\end{pmatrix}\\ &\times\begin{pmatrix}\dfrac{1}{\nu_1+\mu} & 0 & 0\\ 0 & \dfrac{1}{\nu_2+\mu} & 0\\ \dfrac{\nu_1}{(\nu_1+\mu)(\gamma+\mu)} & \dfrac{\nu_2}{(\nu_2+\mu)(\gamma+\mu)} & \dfrac{1}{\gamma+\mu}\end{pmatrix}\\ &=\begin{pmatrix}\dfrac{p\beta\nu_1}{(\nu_1+\mu)(\gamma+\mu)} & \dfrac{p\beta\nu_2}{(\nu_2+\mu)(\gamma+\mu)} & \dfrac{p\beta}{\gamma+\mu}\\ \dfrac{(1-p)\beta\nu_1}{(\nu_1+\mu)(\gamma+\mu)} & \dfrac{(1-p)\beta\nu_2}{(\nu_2+\mu)(\gamma+\mu)} & \dfrac{(1-p)\beta}{\gamma+\mu}\\ 0 & 0 & 0\end{pmatrix}.\end{aligned}$$

As we will show formally in §3, the dominant eigenvalue of this matrix is equal to $\mathcal{R}_0$, where

$$\mathcal{R}_0=\left(\frac{p\nu_1}{\nu_1+\mu}+\frac{(1-p)\nu_2}{\nu_2+\mu}\right)\frac{\beta}{\gamma+\mu}. \tag{2.10}$$

Interlude: From a computational point of view, it is easy to use a mathematical software package to compute $\mathbf{K}_{\mathrm{L}}$ from $\mathbf{T}$ and Σ. We remark, however, that the only cumbersome step, i.e. computing the inverse of Σ, can be easily performed using the biological interpretation of $-\Sigma^{-1}$. In, for example, Diekmann & Heesterbeek (2000, p. 35) it is shown that the element $(-\Sigma^{-1})_{ij}$ is the expected time that an individual who presently has state j will spend in state i during its entire future 'life' (in the epidemiological sense). In the above example this works out as follows. Individuals who are presently in state E_i will spend, on average, an amount of time $1/(\nu_i+\mu)$ in that state. The same individuals will spend on average an amount of time $(\nu_i/(\nu_i+\mu))\times(1/(\gamma+\mu))$ in state I, where the first factor is the probability that an individual actually changes its state from E_i to I, instead of leaving state E_i by dying, and the second factor is the average amount of time an individual who enters state I spends in state I.[3] The individuals in state E_i will spend no time at all in state E_j, with $j\neq i$, leading to zeros for the appropriate elements. Finally, individuals who are presently in state I will spend no time at all in states E_1 and E_2, and will, on average, spend an amount of time $1/(\gamma+\mu)$ in state I. This leads to a full specification of $-\Sigma^{-1}$.

We now proceed with our exposition. The first thing to note is that $\mathbf{T}$ has a special structure: the third row of $\mathbf{T}$ consists of zeros only. Individuals can therefore not be in the third state (in this case state I) immediately after infection. Hence the system has only two states-at-infection: all individuals start their infected life (i.e. are epidemiologically born) in either E_1 or E_2. The NGM is therefore a two-dimensional matrix.

The formal approach to obtaining $\mathbf{K}$ from $\mathbf{K}_{\mathrm{L}}$ is as follows. We pre- and post-multiply $\mathbf{K}_{\mathrm{L}}$ by an auxiliary matrix $\mathbf{E}$ that singles out the rows and columns relevant for the reduced set of states. Specify $\mathbf{E}$ as consisting of unit column vectors $\mathbf{e}_i$, for all i such that the ith row of $\mathbf{T}$ is not identically zero.[4] In other words, create a matrix $\mathbf{E}$ whose columns consist of unit vectors relating to non-zero rows of $\mathbf{T}$ only. In the above case this leads to

$$\mathbf{E}=\begin{pmatrix}1 & 0\\ 0 & 1\\ 0 & 0\end{pmatrix}.$$

To find the NGM we then perform the matrix multiplication

$$\mathbf{K}=\mathbf{E}'\mathbf{K}_{\mathrm{L}}\mathbf{E}=-\mathbf{E}'\mathbf{T}\Sigma^{-1}\mathbf{E}. \tag{2.11}$$

For the example above, we get for the product of $\mathbf{E}'\mathbf{T}$ and $-\Sigma^{-1}\mathbf{E}$

$$\begin{aligned}\mathbf{K}&=\begin{pmatrix}0 & 0 & p\beta\\ 0 & 0 & (1-p)\beta\end{pmatrix}\\ &\times\begin{pmatrix}\dfrac{1}{\nu_1+\mu} & 0\\ 0 & \dfrac{1}{\nu_2+\mu}\\ \dfrac{\nu_1}{(\nu_1+\mu)(\gamma+\mu)} & \dfrac{\nu_2}{(\nu_2+\mu)(\gamma+\mu)}\end{pmatrix}\\ &=\begin{pmatrix}\dfrac{p\beta\nu_1}{(\nu_1+\mu)(\gamma+\mu)} & \dfrac{p\beta\nu_2}{(\nu_2+\mu)(\gamma+\mu)}\\ \dfrac{(1-p)\beta\nu_1}{(\nu_1+\mu)(\gamma+\mu)} & \dfrac{(1-p)\beta\nu_2}{(\nu_2+\mu)(\gamma+\mu)}\end{pmatrix}.\end{aligned}$$

Remember that for a 2×2 matrix the dominant eigenvalue, and hence $\mathcal{R}_0$, can be obtained from the trace and the determinant of the matrix as

$$\begin{aligned}\mathcal{R}_0&=\rho(\mathbf{K})\\ &=\frac{1}{2}\left(\operatorname{trace}(\mathbf{K})+\sqrt{\operatorname{trace}(\mathbf{K})^2-4\det(\mathbf{K})}\right).\end{aligned} \tag{2.12}$$

Note that, in our example, $\det\mathbf{K}=0$, i.e. $\mathbf{K}$ is a singular matrix. Because $\mathbf{K}$ is a 2×2 matrix we can conclude right away that $\mathcal{R}_0=\operatorname{trace}\mathbf{K}$. The resulting expression is as in equation (2.10) above.

Apart from resulting in a simplified expression for $\mathcal{R}_0$ in the two-dimensional case, an NGM with the property that $\det \mathbf{K} = 0$ has the added feature that we can achieve further reduction in dimension of the matrix. We return to this below. First we show that, by epidemiological reasoning directly from the specification of the system, we can obtain the elements of $\mathbf{K}$ from their interpretation without going through the linear algebra. For initial training in this kind of argument it helps to draw a diagram of the system one is studying. For our example above the argument goes as follows. For the element K_{11}, we start with one individual with state-at-infection 1 (i.e. an individual who has just entered state E_1), and determine, by following that individual for the remainder of its infectious life, how many new cases of state-at-infection 1 it is expected to produce. Before the individual can infect, it has to survive the E_1 state and move to the I state. This happens with probability $\nu_1/(\nu_1+\mu)$. While in the I state, the individual is expected to produce new cases at a rate β, for an expected time $1/(\gamma+\mu)$. A fraction p of these will be new cases with state-at-infection 1. Multiplying these factors gives

$$K_{11} = \frac{\nu_1}{\nu_1+\mu}\beta\frac{1}{\gamma+\mu}p.$$

Analogous reasoning gives the expressions for K_{12}, K_{21} and K_{22}.

In this example we saw that $\det \mathbf{K} = 0$. The special feature of the model that causes this is that the states-at-infection are necessarily produced in a fixed proportion.[5] One can then reduce the dimension of the system even further than the reduction from the three-dimensional $\mathbf{K}_{\mathrm{L}}$ to the two-dimensional $\mathbf{K}$. In this example, we need only one state to fully determine $\mathcal{R}_0$, because there is only one state in which individuals can produce new cases, i.e. state I. We will call a state where individuals can produce new cases *state-of-infectiousness*. This argument can be formalized by defining an *NGM with small domain* $\mathbf{K}_{\mathrm{S}}$ for such situations. To see whether the dimension of $\mathbf{K}_{\mathrm{S}}$ is smaller than the dimension of $\mathbf{K}$ we can simply check whether $\det \mathbf{K} = 0$.

To determine $\mathbf{K}_{\mathrm{S}}$ from $\mathbf{K}$, when a reduction is possible, we again examine the transmission matrix $\mathbf{T}$, but instead of only examining the rows we now also examine the columns. For the example above we see that $\mathbf{T}$ has two columns containing only zeros, and only one column that is a non-zero vector. All three columns are therefore multiples of the same vector $\mathbf{C} := (p,\ 1-p,\ 0)'$, the first two columns being zero times this vector, the third column being β times this vector. Similarly, the rows of $\mathbf{T}$ are all multiples of one row vector $\mathbf{R} := (0,\ 0,\ \beta)$, the first row is p times this vector, the second row is $(1-p)$ times this vector, and the third row is zero times this vector. Actually, $\mathbf{R}$ and $\mathbf{C}$ constitute a (multiplicative) decomposition of the transmission matrix $\mathbf{T}$, in the sense that $\mathbf{T} = \mathbf{CR}$, i.e. $\mathbf{T}_{ij} = \mathbf{C}_i\mathbf{R}_j$. We define the NGM with small domain by

$$\mathbf{K}_{\mathrm{S}} = -\mathbf{R}\boldsymbol{\Sigma}^{-1}\mathbf{C}. \tag{2.13}$$

For this example

$$\mathbf{K}_{\mathrm{S}} = -(\,0 \quad 0 \quad \beta\,)\boldsymbol{\Sigma}^{-1}\begin{pmatrix} p \\ 1-p \\ 0 \end{pmatrix} = \left(\frac{p\beta\nu_1}{(\nu_1+\mu)(\gamma+\mu)} + \frac{(1-p)\beta\nu_2}{(\nu_2+\mu)(\gamma+\mu)}\right).$$

The dominant eigenvalue of this 'matrix' equals $\mathcal{R}_0$, as given in equation (2.10).

2.2. *An* SEI *model with two host categories*

To illustrate the power of our approach, we now briefly consider a similar system but allow the difference between categories 1 and 2 in the population to manifest itself in all states. We then distinguish eight states S_1, S_2, E_1, E_2, I_1, I_2, R_1 and R_2, making the system originally six-dimensional. (The sizes of the subpopulations of those that belong to categories 1 and 2 are constant at pN and $(1-p)\,N$, respectively.) The equations for this system are

$$\dot{S}_1 = p\mu N - \beta_{11}\frac{S_1I_1}{N} - \beta_{12}\frac{S_1I_2}{N} - \mu S_1,$$

$$\dot{S}_2 = (1-p)\mu N - \beta_{21}\frac{S_2I_1}{N} - \beta_{22}\frac{S_2I_2}{N} - \mu S_2,$$

$$\dot{E}_1 = \beta_{11}\frac{S_1I_1}{N} + \beta_{12}\frac{S_1I_2}{N} - (\nu_1+\mu)E_1,$$

$$\dot{E}_2 = \beta_{21}\frac{S_2I_1}{N} + \beta_{22}\frac{S_2I_2}{N} - (\nu_2+\mu)E_2,$$

$$\dot{I}_1 = \nu_1E_1 - (\gamma_1+\mu)I_1,$$

$$\dot{I}_2 = \nu_2E_2 - (\gamma_2+\mu)I_2,$$

$$\dot{R}_1 = \gamma_1I_1 - \mu R_1$$

and

$$\dot{R}_2 = \gamma_2I_2 - \mu R_2.$$

Reasoning as in §2.1, we see that there are four infected states in this system, and we restrict ourselves to a four-dimensional infected subsystem. The transmission and transition matrices of the corresponding linearized subsystem are four-dimensional, with

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & p\beta_{11} & p\beta_{12} \\ 0 & 0 & (1-p)\beta_{21} & (1-p)\beta_{22} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} -(\nu_1+\mu) & 0 & 0 & 0 \\ 0 & -(\nu_2+\mu) & 0 & 0 \\ \nu_1 & 0 & -(\gamma_1+\mu) & 0 \\ 0 & \nu_2 & 0 & -(\gamma_2+\mu) \end{pmatrix}.$$

The NGM with large domain $\mathbf{K}_{\mathrm{L}} = -\mathbf{T}\boldsymbol{\Sigma}^{-1}$ is four-dimensional:

$$\mathbf{K}_{\mathrm{L}} = -\mathbf{T}\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \frac{p\beta_{11}\nu_1}{(\nu_1+\mu)(\gamma_1+\mu)} & \frac{p\beta_{12}\nu_2}{(\nu_2+\mu)(\gamma_2+\mu)} \\ \frac{(1-p)\beta_{21}\nu_1}{(\nu_1+\mu)(\gamma_1+\mu)} & \frac{(1-p)\beta_{22}\nu_2}{(\nu_2+\mu)(\gamma_2+\mu)} \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{p\beta_{11}}{\gamma_1+\mu} & \frac{p\beta_{12}}{\gamma_2+\mu} \\ \frac{(1-p)\beta_{21}}{\gamma_1+\mu} & \frac{(1-p)\beta_{22}}{\gamma_2+\mu} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Of the four infected states, there are only two that are states-at-infection. We know this from the model, but we can also see this immediately by looking at $\mathbf{T}$ and noting that two rows consist entirely of zeros. The NGM $\mathbf{K}$ is therefore two-dimensional. The NGM can be found by epidemiological reasoning from the interpretation of its elements in exactly the same way as in §2.1, but replacing the β by the appropriate β_{ij} in K_{ij}. If we use the formal linear algebra approach, we again start by examining the transmission matrix $\mathbf{T}$. The two zero rows are rows 3 and 4. Therefore, the auxiliary matrix $\mathbf{E}$ will have as its columns the first two unit vectors $(1, 0, 0, 0)'$ and $(0, 1, 0, 0)'$:

$$\mathbf{E} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\mathbf{K} = -\mathbf{E}'\mathbf{T}\boldsymbol{\Sigma}^{-1}\mathbf{E} = \begin{pmatrix} \frac{p\beta_{11}\nu_1}{(\nu_1+\mu)(\gamma_1+\mu)} & \frac{p\beta_{12}\nu_2}{(\nu_2+\mu)(\gamma_2+\mu)} \\ \frac{(1-p)\beta_{21}\nu_1}{(\nu_1+\mu)(\gamma_1+\mu)} & \frac{(1-p)\beta_{22}\nu_2}{(\nu_2+\mu)(\gamma_2+\mu)} \end{pmatrix}.$$

We now investigate whether this example allows a further reduction in dimension as in §2.1. We calculate $\det \mathbf{K}$ and establish that it is, in general, not equal to zero. Therefore no further reduction in dimension is possible, unless $\beta_{11}\beta_{22} = \beta_{12}\beta_{21}$. Due to the fact that we have allowed β_{ij} to be different for all combinations, we no longer have that the two states-at-infection occur in a fixed ratio.

For 'completeness' we note that we can regain such a fixed ratio, and the consequent reduction in dimension, in the special case that $\beta_{ij} = a_i b_j$. Here a_i relates to the susceptibility and b_j to the infectivity (so the idea is that the properties of the two individuals involved in a contact that can lead to transmission have an independent influence). This assumption is called separable mixing in Diekmann & Heesterbeek (2000). It leads to $\det \mathbf{K} = 0$, and hence to $R_0 = \operatorname{trace} \mathbf{K}$ because $\mathbf{K}$ is two-dimensional. To proceed formally with this special case via the NGM with small domain, we would write

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & pa_1b_1 & pa_1b_2 \\ 0 & 0 & (1-p)a_2b_1 & (1-p)a_2b_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

To find the NGM with small domain, $\mathbf{K}_{\mathrm{S}}$, we observe that the rows of matrix $\mathbf{T}$ are multiples of the vector $\mathbf{R} = (0, 0, b_1, b_2)$, and the columns are multiples of $\mathbf{C} = (pa_1, (1-p)a_2, 0, 0)'$. Note that $\mathbf{T} = \mathbf{CR}$. We then write

$$\mathbf{K}_{\mathrm{S}} = -(0 \quad 0 \quad b_1 \quad b_2)\boldsymbol{\Sigma}^{-1} \begin{pmatrix} pa_1 \\ (1-p)a_2 \\ 0 \\ 0 \end{pmatrix} = \left(\frac{pa_1b_1\nu_1}{(\nu_1+\mu)(\gamma_1+\mu)} + \frac{(1-p)a_2b_2\nu_2}{(\nu_2+\mu)(\gamma_2+\mu)} \right)$$

and we find

$$R_0 = \frac{p\beta_{11}\nu_1}{(\nu_1+\mu)(\gamma_1+\mu)} + \frac{(1-p)\beta_{22}\nu_2}{(\nu_2+\mu)(\gamma_2+\mu)}.$$

3. RECIPES FOR NGMs

We have now introduced the main concepts of the next-generation approach: the NGM with large domain, the NGM and the NGM with small domain. All three can be derived by simple linear algebra from the basic ingredients $\mathbf{T}$ and $\boldsymbol{\Sigma}$, or by using the epidemiological interpretation of the NGM. In this section we provide recipes for the construction of these matrices by formalizing the steps we have taken in the examples of the previous section. We show in general that $\mathbf{K}$, $\mathbf{K}_{\mathrm{L}}$ and $\mathbf{K}_{\mathrm{S}}$ have the same dominant eigenvalue.

3.1. The NGM with large domain

The NGM with large domain, $\mathbf{K}_{\mathrm{L}}$, is always the matrix with highest dimension. Our starting point is the ODE system that describes the production of new cases and the changes in infected states. We assume that this set of ODEs, the *infection subsystem*, has been written in linearized form. The recipe is as follows.

(i) Decompose[6] the Jacobian matrix of the infection subsystem as $\mathbf{T} + \boldsymbol{\Sigma}$, where $\mathbf{T}$ is the transmission matrix, and $\boldsymbol{\Sigma}$ the transition matrix:
   — $\mathbf{T}$ contains the entries corresponding to transmission events, where an epidemiological birth occurs, and
   — $\boldsymbol{\Sigma}$ contains the entries corresponding to all other changes of state (including death).

(ii) Compute the NGM with large domain as $\mathbf{K}_{\mathrm{L}} = -\mathbf{T}\boldsymbol{\Sigma}^{-1}$.

The ijth entry of $-\boldsymbol{\Sigma}^{-1}$ can be interpreted as the expected time that an individual who presently has infected state j will spend in infected state i (see the

interlude in §2.1). Because the ijth entry of $\mathbf{T}$ is the rate at which an individual in infected state j produces individuals with infected state i, the ijth entry of $\mathbf{K}_{\mathrm{L}} = -\mathbf{T}\boldsymbol{\Sigma}^{-1}$ is the expected number of infected offspring with state i at infection produced throughout its entire future infected life by an individual presently in infected state j. If there are infected states which are not states-at-infection, the matrix $\mathbf{K}_{\mathrm{L}}$ has one or more zero rows. This implies that some of the information contained in $\mathbf{K}_{\mathrm{L}}$ is redundant if we are only interested in the growth or decline of the infected population as we iterate $\mathbf{K}_{\mathrm{L}}$. The NGM $\mathbf{K}$ is the restriction of $\mathbf{K}_{\mathrm{L}}$ to the subset of states-at-infection. Thus this redundancy is removed in $\mathbf{K}$. The interpretation of the entry K_{ij} of $\mathbf{K}$ is the expected number of new infections with state-at-infection i produced by one individual with state-at-infection j.

3.2. The NGM

The NGM, $\mathbf{K}$, has the advantage that it has both a rigorous biological interpretation and excludes irrelevant information. It is usually of lower dimension than $\mathbf{K}_{\mathrm{L}}$. Out of all infected states used for $\mathbf{K}$ we select only those that an infected individual can be in immediately after becoming infected. We call these the states-at-infection, and $\mathbf{K}$ reflects the restriction of the analysis to the states-at-infection. Essentially, as we showed in the examples in §2 (and will also show for the examples in §4) the interpretation allows one to 'compute' $\mathbf{K}$ in a rigorous, but biological, manner. Below, however, we present two linear algebra recipes that allow programming. The second recipe uses the computation of the entire matrix $-\boldsymbol{\Sigma}^{-1}$ and is the easiest when one uses mathematical software to automate the process. The first recipe uses the epidemiological interpretation and demonstrates that one does not need all elements of $-\boldsymbol{\Sigma}^{-1}$ to compute $\mathbf{K}$, some elements will be multiplied by elements of $\mathbf{T}$ that are zero and therefore do not contribute. In fact, the second recipe is a programmable version of the first.

The first recipe is as follows.

(i) Identify, see §3.1, the transmission matrix $\mathbf{T}$, and the transition matrix $\boldsymbol{\Sigma}$.
(ii) Identify the *states-at-infection*. State j is a state-at-infection if and only if there is at least one non-zero element in the jth row of matrix $\mathbf{T}$.
(iii) Identify the *states-of-infectiousness*. State ℓ is a state-of-infectiousness if and only if there is at least one non-zero element in the ℓth column of $\mathbf{T}$.
(iv) Compute an auxiliary matrix $\mathbf{A}$ which has elements $\mathbf{A}_{\ell j} := -\,(\boldsymbol{\Sigma}^{-1})_{\ell j}$ for all ℓj combinations where j is a state-at-infection and ℓ is a state-of-infectiousness, and for which all other elements are zero.
(v) Define $\mathbf{K}_{ij} = (\mathbf{TA})_{ij}$ for all combinations with i and j both states-at-infection.

The second recipe is as follows.

(i) Determine if the number of states-at-infection is less than the dimension of the infection subsystem.
— If $\mathbf{T}$ has no rows consisting entirely of zeros, then $\mathbf{K} = \mathbf{K}_{\mathrm{L}}$ and proceed with step (ii) in §3.1.
— If $\mathbf{T}$ has one or more rows consisting entirely of zeros, then $\mathbf{K} \neq \mathbf{K}_{\mathrm{L}}$ and proceed as below.
(ii) Identify the auxiliary matrix $\mathbf{E}$ as follows:
— The matrix $\mathbf{E}$ has the same number of rows as $\mathbf{T}$.
— There is one column of $\mathbf{E}$ for each non-zero row of $\mathbf{T}$, and hence for each state-at-infection. That column of $\mathbf{E}$ has a one in the row that corresponds to the non-zero row of $\mathbf{T}$ (state-at-infection), and zeros elsewhere.
(iii) Compute the NGM, $\mathbf{K} = -\mathbf{E}'\,\mathbf{T}\boldsymbol{\Sigma}^{-1}\,\mathbf{E}$.

By definition the basic reproduction number is the largest eigenvalue of the NGM, $R_0 = \rho(\mathbf{K})$. We now show that the NGM and the NGM with large domain have the same non-zero eigenvalues. Let $\mathbf{v}$ be an eigenvector of $\mathbf{K}$ with corresponding eigenvalue λ. Then $\mathbf{Kv} = -\mathbf{E}'\mathbf{T}\boldsymbol{\Sigma}^{-1}\mathbf{Ev} = \lambda\mathbf{v}$. Multiply this identity by $\mathbf{E}$ to get $-\mathbf{EE}'\mathbf{T}\boldsymbol{\Sigma}^{-1}\mathbf{Ev} = \lambda\mathbf{Ev}$. But $\mathbf{EE}'\mathbf{T} = \mathbf{T}$, so $\mathbf{Ev}$ is an eigenvector of $\mathbf{K}_{\mathrm{L}}$ with corresponding eigenvalue λ, and the non-zero eigenvalues of $\mathbf{K}$ and $\mathbf{K}_{\mathrm{L}}$ are the same. (Note that it is impossible that $\mathbf{Ev} = 0$ because this would imply that $\lambda\mathbf{v} = \mathbf{Kv} = 0$, hence $\mathbf{v} = 0$ as $\lambda \neq 0$.)

3.3. The NGM with small domain

The NGM with small domain, $\mathbf{K}_{\mathrm{S}}$, has the lowest dimension of the three types of NGM discussed. In many cases, however, it will be equal to $\mathbf{K}$. If $\det \mathbf{K} = 0$, the NGM with small domain is different from $\mathbf{K}$. This will certainly be the case if there are fewer states-of-infectiousness than states-at-infection, as in the example in §2.1 above (and in the example in §4.2). Indeed, in that case it makes perfect sense to define a matrix $\mathbf{K}_{\mathrm{S}}$ with elements $\mathbf{K}_{\mathbf{S}_{ij}} = -(\mathbf{T}\boldsymbol{\Sigma}^{-1})_{ij}$, with both i and j restricted to states-of-infectiousness. It simply means that we focus our bookkeeping on individuals who have just entered a state-of-infectiousness, and compute how many of their epidemiological offspring will enter, on average, the various states-of-infectiousness. In other words, we base our bookkeeping not on being born, but on the later phase in the 'epidemiological life' where the individual starts to reproduce.[7]

As the example presented in §2.2 shows, there may be other reasons why $\det \mathbf{K} = 0$. We now give a general recipe to derive $\mathbf{K}_{\mathrm{S}}$ from $\mathbf{K}$ (in a manner that works whatever the reason is that $\det \mathbf{K} = 0$). The recipe is as follows.

(i) Follow the recipe in §3.2 to determine $\mathbf{K}$.
(ii) Determine whether $\det \mathbf{K} = 0$.[8]
— If $\det \mathbf{K} \neq 0$ then no further reduction is possible and $\mathbf{K}_{\mathrm{S}} = \mathbf{K}$.
— If $\det \mathbf{K} = 0$ proceed as below.
(iii) Define a matrix $\mathbf{R}$, whose rows are linearly independent vectors spanning the rows of $\mathbf{T}$, and a matrix $\mathbf{C}$, whose columns are linearly independent vectors spanning the columns of $\mathbf{T}$. Scale the matrices so that $\mathbf{T} = \mathbf{CR}$.
(iv) Compute the NGM with small domain, $\mathbf{K}_{\mathrm{S}} = -\mathbf{R}\boldsymbol{\Sigma}^{-1}\,\mathbf{C}$.

As a side remark we now explain that one can derive $\mathbf{K}_{\mathrm{S}}$ from $\mathbf{K}$ in more or less the same way as we derived

$\mathbf{K}$ from $\mathbf{K}_{\mathrm{L}}$. When deriving $\mathbf{K}$, we consider 'pure' states-at-infection and represent these in the columns of $\mathbf{E}$. A more general point of view considers 'mixed' states-at-infection, by which we mean a probability distribution for state-at-infection represented by a column with non-negative elements that sum to one (and with zero elements, of course, at positions that do not correspond to states-at-infection). By replacing $\mathbf{E}$ by a matrix consisting of such probability vectors, one may derive $\mathbf{K}_{\mathrm{S}}$ directly from $-\mathbf{T}\boldsymbol{\Sigma}^{-1}$, following the recipe in §3.2.

We now show that the NGM with small domain and the NGM with large domain have the same non-zero eigenvalues. Let $\mathbf{v}$ be an eigenvector of $\mathbf{K}_{\mathrm{S}}$ with corresponding eigenvalue λ. Then $\mathbf{K}_{\mathrm{S}}\mathbf{v} = -\mathbf{R}\boldsymbol{\Sigma}^{-1}\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$. Multiply this identity by $\mathbf{C}$ to get $-\mathbf{C}\mathbf{R}\boldsymbol{\Sigma}^{-1} \times \mathbf{C}\mathbf{v} = \lambda\mathbf{C}\mathbf{v}$. But $\mathbf{C}\mathbf{R} = \mathbf{T}$, so $\mathbf{C}\mathbf{v}$ is an eigenvector of $\mathbf{K}_{\mathrm{L}}$ with corresponding eigenvalue λ. As the matrices $\mathbf{K}$, $\mathbf{K}_{\mathrm{L}}$ and $\mathbf{K}_{\mathrm{S}}$ have the same rank, we have established that they have the same non-zero spectrum and hence the same dominant eigenvalue.

4. EXAMPLES

To illustrate our method further, we present three more examples, each highlighting special difficulties one might encounter. For the first example we analyse a sexually transmitted infection of *SEI* type, which we then extend by adding vertical transmission of infection. The final example is taken from the literature and based on a model for the transmission of bovine viral diarrhoea. For each example we start with the infection subsystem.

4.1. A sexual transmission SEI *model*

Consider a purely heterosexually transmitted infectious disease. If the numbers of exposed and infectious females are E_1 and I_1, and the numbers of exposed and infectious males are E_2 and I_2 respectively, then we assume that

$$\dot{E}_1 = \beta_1 S_1 \frac{I_2}{N_2} - (\nu_1 + \mu)E_1,$$

$$\dot{I}_1 = \nu_1 E_1 - (\gamma_1 + \mu)I_1,$$

$$\dot{E}_2 = \beta_2 S_2 \frac{I_1}{N_1} - (\nu_2 + \mu)E_2$$

and

$$\dot{I}_2 = \nu_2 E_2 - (\gamma_2 + \mu)I_2,$$

where N_1 and N_2 are the sizes of the subpopulations of females and males, respectively. To construct the NGM $\mathbf{K}$, observe that a newly infected male (in the E_2 state or with state-at-infection E_2) has a probability $\nu_2/(\nu_2 + \mu)$ of entering the I_2 state, and would then infect females at a rate $\beta_1 N_1/N_2$ over a period of $1/(\gamma_2 + \mu)$ time units. Hence the entry in row one column two. A similar argument specifies the entry in row two column one. We have deduced that

$$\mathbf{K} = \begin{pmatrix} 0 & \dfrac{\nu_2\beta_1 N_1}{(\nu_2+\mu)(\gamma_2+\mu)N_2} \\ \dfrac{\nu_1\beta_2 N_2}{(\nu_1+\mu)(\gamma_1+\mu)N_1} & 0 \end{pmatrix}. \tag{4.1}$$

Hence we obtain the expression for $\mathcal{R}_0$ directly from the formula (2.12) with trace $\mathbf{K} = 0$:

$$R_0 = \rho(\mathbf{K}) = \sqrt{\frac{\nu_1\nu_2\beta_1\beta_2}{(\nu_1+\mu)(\nu_2+\mu)(\gamma_1+\mu)(\gamma_2+\mu)}}. \tag{4.2}$$

This example illustrates how easy it is to write down the NGM directly from epidemiological reasoning. As the NGM is two-dimensional it is then straightforward to compute $\mathcal{R}_0$.

For the more laborious way of using the recipe one proceeds as follows. Specify, from the infection subsystem, the transmission matrix as

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & 0 & \beta_1\dfrac{N_1}{N_2} \\ 0 & 0 & 0 & 0 \\ 0 & \beta_2\dfrac{N_2}{N_1} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and the transition matrix as

$$\boldsymbol{\Sigma} = \begin{pmatrix} -(\nu_1+\mu) & 0 & 0 & 0 \\ \nu_1 & -(\gamma_1+\mu) & 0 & 0 \\ 0 & 0 & -(\nu_2+\mu) & 0 \\ 0 & 0 & \nu_2 & -(\gamma_2+\mu) \end{pmatrix}.$$

Then calculate

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} -\dfrac{1}{\nu_1+\mu} & 0 & 0 & 0 \\ -\dfrac{\nu_1}{(\nu_1+\mu)(\gamma_1+\mu)} & -\dfrac{1}{\gamma_1+\mu} & 0 & 0 \\ 0 & 0 & -\dfrac{1}{\nu_2+\mu} & 0 \\ 0 & 0 & -\dfrac{\nu_2}{(\nu_2+\mu)(\gamma_2+\mu)} & -\dfrac{1}{\gamma_2+\mu} \end{pmatrix}.$$

The NGM with large domain is given by $\mathbf{K}_{\mathrm{L}} = -\mathbf{T}\boldsymbol{\Sigma}^{-1}$ and is four-dimensional. Using the second recipe in §3.2, we observe that the matrix $\mathbf{T}$ has only two non-zero rows, the first and third, so the auxiliary matrix $\mathbf{E}$ is given by

$$\mathbf{E} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

and the formula $\mathbf{K} = -\mathbf{E}'\mathbf{T}\boldsymbol{\Sigma}^{-1}\mathbf{E}$ then leads to equation (4.1) above. Note that $\det\mathbf{K} \neq 0$, so no further reduction of dimension ($\mathbf{K}$ to $\mathbf{K}_{\mathrm{S}}$) is possible.

4.2. A model for a sexually transmitted infection with vertical transmission

Now consider an *SI* model for a heterosexually transmitted infectious disease that may also be transmitted vertically. As new-born individuals are not immediately sexually active, we take J_1 and J_2 to be the numbers of infected juvenile females and males, and I_1 and I_2 to be

the numbers of infected adult females and males, respectively. We assume that both the length of the pre-sexual period and the length of the infectious period are large compared to the latency period, so we neglect the latter. We also assume that the sex ratio of offspring is one-to-one (a logical consequence is that $N_1 = N_2$ if the *per capita* death rates are equal, but we shall keep the quasi-generality of allowing these numbers to be different). We are thus led to consider the following infected subsystem:

$$\begin{aligned}\dot{J}_1 &= p\mu I_1 - (\nu_1 + \mu)J_1,\\ \dot{I}_1 &= \nu_1 J_1 + \beta_1 S_1 \frac{I_2}{N_2} - (\gamma_1 + \mu)I_1,\\ \dot{J}_2 &= p\mu I_1 - (\nu_2 + \mu)J_2\end{aligned}$$

and

$$\dot{I}_2 = \nu_2 J_2 + \beta_2 S_2 \frac{I_1}{N_1} - (\gamma_2 + \mu)I_2,$$

where p denotes the probability that a vertical transmission takes place when offspring is produced.

There are four *states-at-infection*: vertically infected females J_1; horizontally infected females (included in I_1); vertically infected males J_2; and horizontally infected males (included in I_2). A horizontally infected female is initially in the I_1 state. She produces vertically infected females and males at rate $p\mu$ and horizontally infects males at rate $\beta_2 N_2/N_1$, all for a period of, on average, $1/(\gamma_1 + \mu)$ time units. Hence the second column of $\mathbf{K}$, specified in equation (4.3). A vertically infected female enters the I_1 state with probability $\nu_1/(\nu_1 + \mu)$, hence the first column of $\mathbf{K}$ is just a multiple of the second. A horizontally infected male is initially in the I_2 state, and horizontally infects females at the rate $\beta_1 N_1/N_2$ for a period of $1/(\gamma_2 + \mu)$ time units, hence the K_{24} entry. This is the only way that a male transmits the infection, so the other entries in the fourth column of $\mathbf{K}$ are zero. Finally, a vertically infected male enters the I_2 state with probability $\nu_2/(\nu_2 + \mu)$, and the third column of $\mathbf{K}$ is a multiple of the fourth. Note that all of these expressions concern a fully susceptible population. The NGM for this model is

$$\mathbf{K} = \begin{pmatrix} \dfrac{\nu_1 p\mu}{(\nu_1+\mu)(\gamma_1+\mu)} & \dfrac{p\mu}{\gamma_1+\mu} \\ 0 & 0 \\ \dfrac{\nu_1 p\mu}{(\nu_1+\mu)(\gamma_1+\mu)} & \dfrac{p\mu}{\gamma_1+\mu} \\ \dfrac{\nu_1\beta_2 N_2}{(\nu_1+\mu)(\gamma_1+\mu)N_1} & \dfrac{\beta_2 N_2}{(\gamma_1+\mu)N_1} \\ 0 & 0 \\ \dfrac{\nu_2\beta_1 N_1}{(\nu_2+\mu)(\gamma_2+\mu)N_2} & \dfrac{\beta_1 N_1}{(\gamma_2+\mu)N_2} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}. \qquad (4.3)$$

The NGM is four-dimensional, but of rank two and hence has two zero eigenvalues.

We have shown how the NGM is easily constructed by epidemiological reasoning. Alternatively, we may proceed using the linear algebra recipe. We observe that the transmission and transition matrices are, respectively,

$$\mathbf{T} = \begin{pmatrix} 0 & p\mu & 0 & 0 \\ 0 & 0 & 0 & \beta_1 \dfrac{N_1}{N_2} \\ 0 & p\mu & 0 & 0 \\ 0 & \beta_2 \dfrac{N_2}{N_1} & 0 & 0 \end{pmatrix}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} -(\nu_1+\mu) & 0 & 0 & 0 \\ \nu_1 & -(\gamma_1+\mu) & 0 & 0 \\ 0 & 0 & -(\nu_2+\mu) & 0 \\ 0 & 0 & \nu_2 & -(\gamma_2+\mu) \end{pmatrix}.$$

As $\mathbf{T}$ has entries in all four rows, each infected state is, in this example, a state-at-infection. Because there are no zero rows the matrix $\mathbf{E}$ consists of all four unit vectors and therefore equals the identity matrix. Hence for this example $\mathbf{K} = \mathbf{K}_{\mathrm{L}} = -\mathbf{T}\mathbf{\Sigma}^{-1}$.

Note that, from $\mathbf{T}$, it is easily seen that among the four states-at-infection, there are only two that are also states-of-infectiousness: only the second and fourth columns of $\mathbf{T}$ contain at least one non-zero element. The columns correspond to state I_1 and I_2, which can, of course, also be gleaned from the biological interpretation of the four states.

So $\det \mathbf{K} = 0$, and reduction to an NMG with small domain is possible. The matrix $\mathbf{K}_{\mathrm{S}}$ has eigenvalues equal to the two non-zero eigenvalues of $\mathbf{K}$. To formally construct the matrix $\mathbf{K}_{\mathrm{S}}$ we observe that the rows of matrix $\mathbf{T}$ are spanned by the vectors $(0,1,0,0)$ and $(0,0,0,1)$; and the columns are spanned by $(p\mu, 0, p\mu, \beta_2 N_2/N_1)'$ and $(0, \beta_1 N_1/N_2, 0, 0)'$. We then define matrices

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{C} = \begin{pmatrix} p\mu & 0 \\ 0 & \dfrac{\beta_1 N_1}{N_2} \\ p\mu & 0 \\ \dfrac{\beta_2 N_2}{N_1} & 0 \end{pmatrix}$$

and write

$$\begin{aligned}\mathbf{K}_{\mathrm{S}} &= -\mathbf{R}\,\mathbf{\Sigma}^{-1}\,\mathbf{C}\\ &= \begin{pmatrix} \dfrac{\nu_1 p\mu}{(\nu_1+\mu)(\gamma_1+\mu)} & \dfrac{\beta_1 N_1}{(\gamma_2+\mu)N_2} \\ \dfrac{\beta_2 N_2}{(\gamma_1+\mu)N_1} + \dfrac{\nu_2 p\mu}{(\nu_2+\mu)(\gamma_1+\mu)} & 0 \end{pmatrix}.\end{aligned}$$

The basic reproduction number can be obtained easily from the formula for 2×2 matrices (equation (2.12)) applied to $\mathbf{K}_{\mathrm{S}}$:

$$R_0 = \rho(\mathbf{K}) = \rho(\mathbf{K}_{\mathrm{S}}) = \frac{\nu_1 p\mu}{2(\nu_1+\mu)(\gamma_1+\mu)} + \sqrt{\left(\frac{\nu_1 p\mu}{2(\nu_1+\mu)(\gamma_1+\mu)}\right)^2 + \frac{\nu_2 p\mu\beta_1 N_1 + \beta_1\beta_2(\nu_2+\mu)N_2}{(\nu_2+\mu)(\gamma_1+\mu)(\gamma_2+\mu)N_2}}.$$

4.3. A model for bovine viral diarrhoea

A modified *SEIR* model for bovine viral diarrhoea was described by Cherry *et al.* (1998). The system has both horizontal and vertical transmissions. Horizontally infected animals can be in E, I and R states. Animals that have been pregnant for less than 150 days when becoming infected may, following recovery into one particular immune state Z of several possible immune states, give birth to an infected calf. These offspring are classified as *persistently infected* (P state): they transmit infection, give birth at a lower rate and die at a higher rate than cattle that were infected by the horizontal route. Let the constant p_1 be the probability that an infected animal enters the immune state Z upon recovery, let p_2 be the probability that an infected foetus survives to enter the herd, $1/\alpha$ be the average time spent carrying an infected foetus, and a and b be the reduction in birth rate and increase in death rate of persistently infected animals, respectively.

With a change in notation from Cherry *et al.* the model is described by

$$\begin{aligned}\dot{E} &= (\beta_1 I + \beta_2 P)S - (\nu+\mu)E,\\ \dot{I} &= \nu E - (\gamma+\mu)I,\\ \dot{Z} &= p_1\gamma I - (\alpha+\mu)Z\end{aligned}$$

and

$$\dot{P} = p_2\alpha Z + (\mu - a)P - (\mu + b)P,$$

where, as before, we restrict ourselves to the infected subsystem. Note, however, that there is a difference with the previous examples that included a recovered state. In the previous examples the R state did not occur in the equations for the infected states and could therefore be ignored for the construction of the NGM and the calculation of $\mathcal{R}_0$. Individuals in an R state do not give rise to new infections, so R is considered to be a non-infected state. In this particular example this is still true as far as horizontal transmission is concerned and for all immune states in the model (not shown here) other than state Z. It is, however, not true that recovered individuals cannot produce new infections in this model, because vertical transmission occurs from immune state Z. Calves are born after the mother has recovered from the infection, and therefore recovered mothers in state Z can give rise to new infections through birth. There are therefore four infected states. The transmission and transition matrices are

$$\mathbf{T} = \begin{pmatrix} 0 & \beta_1 & 0 & \beta_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & p_2\alpha & \mu - a \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} -(\nu+\mu) & 0 & 0 & 0 \\ \nu & -(\gamma+\mu) & 0 & 0 \\ 0 & p_1\gamma & -(\alpha+\mu) & 0 \\ 0 & 0 & 0 & -(\mu+b) \end{pmatrix}.$$

We omit the computation of the four-dimensional $\mathbf{K}_{\mathrm{L}}$. Note that $\mathbf{T}$ has non-zero elements in two rows only, hence $\mathbf{K}$ is two-dimensional. The two states-at-infection are the horizontally infected E state and the vertically (and persistently) infected P state. Define

$$\mathbf{E} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then the NGM is given by

$$\begin{aligned}\mathbf{K} &= -\mathbf{E}'\mathbf{T}\,\Sigma^{-1}\,\mathbf{E}\\ &= \begin{pmatrix} \dfrac{\nu\beta_1}{(\nu+\mu)(\gamma+\mu)} & \dfrac{\beta_2}{\mu+b} \\ \dfrac{p_2\alpha\nu p_1\gamma}{(\nu+\mu)(\gamma+\mu)(\alpha+\mu)} & \dfrac{\mu-a}{\mu+b} \end{pmatrix}.\end{aligned}$$

The basic reproduction number is then calculated easily from $\mathbf{K}$ using equation (2.12).

The NGM $\mathbf{K}$ can, of course, also be constructed directly from epidemiological considerations as follows. A proportion $\nu/(\nu+\mu)$ of those horizontally infected become infectious, and infect others at the rate β_1 for, on average, $1/(\gamma+\mu)$ time units. A proportion $p_1\gamma/(\alpha+\mu)$ of these enter the Z state, and give birth to persistently infected calves at rate $p_2\alpha$ for, on average, $1/(\mu+b)$ time units. Hence the first column of $\mathbf{K}$ is obtained. To construct the second column, observe that those in the persistently infected state infect others horizontally at rate β_2, and give birth to persistently infected calves at rate $\mu - a$ for $1/(\mu+b)$ time units. So we see, once again, that a simple epidemiological argument may be used to directly construct the NGM $\mathbf{K}$.

5. CONCLUSION

In epidemic models, individuals can typically be in a number of different states, reflecting both differences in traits and differences in infection status. From the states that apply to infected individuals, we single out those states that individuals can be in *immediately* after they have been infected. We call these states-at-infection. They play a special role in the definition and calculation of $\mathcal{R}_0$ as the dominant eigenvalue of the NGM associated with the epidemic system. The NGM has an appealing epidemiological interpretation because its components may be regarded as $\mathcal{R}_0$-like quantities. We have provided a recipe for the construction of the NGM for general compartmental epidemic models, exploiting also that there may be only a few states-of-infectiousness in a given system. The recipe may be implemented easily in commonly used mathematical software.

We have in fact given three recipes because we have identified three different NGMs and have clarified the relationships between them. This is useful because some researchers have been confused when trying to reconcile an existing algorithmic linear algebra approach (Diekmann & Heesterbeek 2000, pp. 105–107; Van den Driessche & Watmough 2002) with the original approach using the epidemiological interpretation. We show that the reason is that the approaches lead to

two different matrices, which we now call the NGM and the NGM with large domain. Both of these matrices have $\mathcal{R}_0$ as their dominant eigenvalue, the difference lies in the set of individual states that the matrices reflect. We have provided easy algorithms for the construction of the matrices. Both algorithms start by identifying transmission and transition matrices from the linearization of the compartmental model near the infection-free steady state: the transmission matrix describes the production of new infections, and the transition matrix describes changes of infected states (including removal by death or recovery). The NGM with large domain is obtained by a direct construction using these two matrices. By identifying the subset of epidemiological states-at-infection, which is easily done by examining the transmission matrix, we use the second recipe to find the NGM proper. It is often of lower dimension than the NGM with large domain, leading to a simpler calculation of $\mathcal{R}_0$. Sometimes it is possible to construct the NGM with small domain. This matrix may have a less readily understandable interpretation in terms of the epidemiology, but has the advantage of a lower dimension.

Although we present three mathematical recipes, we encourage the construction of the NGM from epidemiological reasoning. This is straightforward and maintains the connection between the mathematics and the biology, and especially gives the user a fuller understanding of the interpretation of the results.

O.D. acknowledges support by the CRM, Universitat Autonoma, Barcelona, and by the Hokkaido University, Sapporo, during his sabbatical leave. J.A.P.H. was supported by grant 918.56.620 of ZonMw/NWO. M.G.R. received support from the Marsden Fund under contract MAU0809. He also acknowledges Massey University and the University of Utrecht for supporting his sabbatical period of research in Utrecht. The authors are very grateful to Gary Smith of the University of Pennsylvania (and RLH) for persistently asking for more clarification over the years and for beta-testing our exposition, to Barbara Boldin for pointing out some errors and to Hans Schneider for critically reviewing appendix A.

APPENDIX A. A PROOF THAT $\mathcal{R}_0$ GOVERNS THE STABILITY OF THE INFECTION-FREE STEADY STATE

Before formulating the key hypotheses concerning $\mathbf{T}$ and $\mathbf{\Sigma}$ we introduce some notation. For a square matrix $\mathbf{A}$ we denote by $s(\mathbf{A})$ the *spectral bound* and by $\rho(\mathbf{A})$ the *spectral radius*:

$$s(\mathbf{A}) := \sup\{\mathrm{Re}(\lambda) : \lambda \in \sigma(\mathbf{A})\}$$

and

$$\rho(\mathbf{A}) := \sup\{|\lambda| : \lambda \in \sigma(\mathbf{A})\},$$

where $\sigma(\mathbf{A})$ denotes the *spectrum* of $\mathbf{A}$, that is the set of eigenvalues. All matrices that we consider have real entries. As customary, we call a non-zero matrix $\mathbf{A}$ *positive* if all entries are non-negative; and *positive-off-diagonal* if all entries are non-negative except possibly those on the diagonal. The following holds if $\mathbf{A}$ is a positive-off-diagonal matrix: $s(\mathbf{A}) < 0$ if and only if $\mathbf{A}$ is invertible and $-\mathbf{A}^{-1}$ is a positive matrix (for a proof see, for example, lemma 6.12 in Diekmann & Heesterbeek (2000)).

In the following we assume that $\mathbf{T}$ is a positive matrix, and that $\mathbf{\Sigma}$ is a positive-off-diagonal matrix with $s(\mathbf{\Sigma}) < 0$, hence $-\mathbf{\Sigma}^{-1}$ is a positive matrix. These assumptions reflect the biological meaning of both matrices; the condition $s(\mathbf{\Sigma}) < 0$ reflects that one cannot remain (potentially) infectious for ever.

For the proof it is convenient to take the NGM with large domain $\mathbf{K}_\mathrm{L}$ as our starting point; the equivalence of the spectral radius of $\mathbf{K}_\mathrm{L}$ and $\mathbf{K}$, as shown in §3, then confirms the result for the NGM $\mathbf{K}$. The basic reproduction number $\mathcal{R}_0$ is defined by

$$\mathcal{R}_0 = \rho(\mathbf{K}) = \rho(\mathbf{K}_\mathrm{L}) = \rho(-\mathbf{T}\mathbf{\Sigma}^{-1}).$$

The stability of the zero steady state of the linear system

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = (\mathbf{T} + \mathbf{\Sigma})\mathbf{x}$$

is determined by the sign of the Malthusian parameter r, which is defined as

$$r = s(\mathbf{T} + \mathbf{\Sigma}).$$

This criterion extends to the nonlinear system by the principle of linearized stability if, in addition, the demographic dynamics make the infection-free steady state stable in the invariant subspace corresponding to the absence of the infectious agent. The key result of this appendix is the following.

Theorem A.1. *Let* $\mathbf{T}$ *be a positive matrix and let* $\mathbf{\Sigma}$ *be a positive off-diagonal matrix with* $s(\mathbf{\Sigma}) < 0$. *Let* $\mathcal{R}_0 = \rho(-\mathbf{T}\mathbf{\Sigma}^{-1})$ *and* $r = s(\mathbf{T} + \mathbf{\Sigma})$. *Then the following equality holds*:[9]

$$\mathrm{sign}\,(r) = \mathrm{sign}(\mathcal{R}_0 - 1).$$

We first prove the result under the extra assumptions that $\mathbf{T} + \mathbf{\Sigma}$ is irreducible and $\mathcal{R}_0 > 0$, and then employ an *approximation and continuity* argument to establish the result in general. The proof is based on ideas in Li & Schneider (2002), who addressed a similar problem in population dynamics in a discrete-time setting. In Van den Driessche & Watmough (2002) a proof is presented in terms of M-matrices, and we refer to Thieme (2009) for the analogous result for the infinite dimensional case.

Lemma A.2. *If* $\mathcal{R}_0 > 0$ *then* $s(\mathcal{R}_0^{-1}\,\mathbf{T} + \mathbf{\Sigma}) = 0$.

Proof. First assume that $\mathbf{T} + \mathbf{\Sigma}$ is irreducible. Let $\mathbf{v}$ be the non-negative left eigenvector of $\mathbf{K}_\mathrm{L} = -\mathbf{T}\mathbf{\Sigma}^{-1}$ corresponding to the eigenvalue $\mathcal{R}_0$. Hence $\mathbf{v}\mathbf{K}_\mathbf{L} = \mathcal{R}_0\mathbf{v}$ which can be rearranged to obtain

$$\mathbf{v}\big(\mathcal{R}_0^{-1}\mathbf{T} + \mathbf{\Sigma}\big) = 0. \tag{A 1}$$

The irreducibility of $\mathbf{T} + \mathbf{\Sigma}$ implies that $\mathcal{R}_0^{-1}\,\mathbf{T} + \mathbf{\Sigma}$ is irreducible. By adding a large positive multiple of the identity, $k\mathbf{I}$, to $\mathcal{R}_0^{-1}\,\mathbf{T} + \mathbf{\Sigma}$, we obtain a positive irreducible matrix, and since $\mathbf{v}$ is non-negative it must be the eigenvector corresponding to the spectral radius of that

matrix. It follows that all the other eigenvalues have smaller real parts. By subtracting $k\mathbf{I}$ again all eigenvalues shift to the left in the complex plane, but the order relation between their real parts remains intact. Hence we conclude from equation (A 1) that $s(\mathcal{R}_0^{-1}\mathbf{T}+\mathbf{\Sigma})=0$.

Next consider the case that $\mathbf{T}+\mathbf{\Sigma}$ is reducible. Regard the irreducible matrix $\mathbf{T}+\epsilon\mathbf{1}+\mathbf{\Sigma}$, where $\mathbf{1}$ is the matrix with all entries equal to one. Denote the spectral radius of the matrix $-(\mathbf{T}+\epsilon\mathbf{1})\mathbf{\Sigma}^{-1}$ by ρ_ϵ. For $\epsilon\downarrow 0$, we have that $\rho_\epsilon\to\mathcal{R}_0$ and hence $\rho_\epsilon>0$ for ϵ small. So, by the above proof for the irreducible case, $s((\mathbf{T}+\epsilon\mathbf{1})/\rho_\epsilon+\mathbf{\Sigma})=0$. Finally, for $\epsilon\downarrow 0$ we have, as noted above, that $\rho\epsilon\to\mathcal{R}_0$, and hence $s(\mathbf{T}/\mathcal{R}_0+\mathbf{\Sigma})=\lim_{\epsilon\to 0}s((\mathbf{T}+\epsilon\mathbf{1})/\rho_\epsilon+\mathbf{\Sigma})=0$. ■

Lemma A.3. *If* $\mathbf{T}+\mathbf{\Sigma}$ *is irreducible then*

$$y\mapsto s(y^{-1}\mathbf{T}+\mathbf{\Sigma})$$

is strictly monotone decreasing.

Proof. We first add $k\mathbf{I}$ to $\mathbf{T}+\mathbf{\Sigma}$ for some k large enough to obtain a positive matrix. The spectral radius of an irreducible positive matrix strictly decreases (increases) if any entry of that matrix decreases (increases) (see theorem 2.1 in Li & Schneider (2002) and references therein). Hence the spectral radius of $y^{-1}\mathbf{T}+\mathbf{\Sigma}+k\mathbf{I}$ is a monotone function of y. For a positive matrix the spectral radius is equal to the spectral bound, and it remains equal to the spectral bound as the spectrum shifts to the left when we subtract $k\mathbf{I}$. ■

Lemma A.4. *If* $\mathbf{T}+\mathbf{\Sigma}$ *is irreducible and* $\mathcal{R}_0>0$ *then* sign $(r)=\mathrm{sign}(\mathcal{R}_0-1)$.

Proof. If $\mathcal{R}_0>1$ then (by lemma A.3) $s(\mathbf{T}+\mathbf{\Sigma})>s(\mathcal{R}_0^{-1}\mathbf{T}+\mathbf{\Sigma})$, but (by lemma A.2) $s(\mathcal{R}_0^{-1}\mathbf{T}+\mathbf{\Sigma})=0$, hence $r=s(\mathbf{T}+\mathbf{\Sigma})>0$. If $\mathcal{R}_0=1$ then (by lemma A.2) $r=s(\mathbf{T}+\mathbf{\Sigma})=0$. If $\mathcal{R}_0<1$ then (by lemma A.3) $s(\mathbf{T}+\mathbf{\Sigma})<s(\mathcal{R}_0^{-1}\mathbf{T}+\mathbf{\Sigma})$, by lemma A.2 $s(\mathcal{R}_0^{-1}\mathbf{T}+\mathbf{\Sigma})=0$, hence $r=s(\mathbf{T}+\mathbf{\Sigma})<0$. ■

Lemma A.5. *If* $s(\mathbf{T}+\mathbf{\Sigma})=0$ *then* $\mathcal{R}_0\geq 1$.

Proof. By the shifting argument used above, it follows that $s(\mathbf{T}+\mathbf{\Sigma})$ is an eigenvalue of $\mathbf{T}+\mathbf{\Sigma}$. Let $\mathbf{u}\neq 0$ be a vector such that $(\mathbf{T}+\mathbf{\Sigma})\mathbf{u}=0$, and define $\mathbf{v}=\mathbf{\Sigma}\mathbf{u}$. As $\mathbf{\Sigma}$ is invertible, $\mathbf{v}\neq 0$. Moreover, $(\mathbf{T}\mathbf{\Sigma}^{-1}+\mathbf{I})\mathbf{v}=(\mathbf{T}+\mathbf{\Sigma})\mathbf{u}=0$, hence $\mathbf{K}_\mathrm{L}=-\mathbf{T}\mathbf{\Sigma}^{-1}$ has a unit eigenvalue and the spectral radius of $\mathbf{K}_\mathrm{L}$ must be greater than or equal to one. ■

Lemma A.6. *If* $s(\mathbf{T}+\mathbf{\Sigma})=0$ *then* $\mathcal{R}_0=1$.

Proof. Below we approximate $\mathbf{K}_\mathrm{L}=-\mathbf{T}\mathbf{\Sigma}^{-1}$ with a continuous family of matrices, parametrized by ϵ, that have spectral radius less than or equal to one for $\epsilon>0$, and which converge to $\mathbf{K}_\mathrm{L}$ as $\epsilon\downarrow 0$. It follows that $\mathcal{R}_0\leq 1$. Because from lemma A.5 it follows that $\mathcal{R}_0\geq 1$, we conclude that $\mathcal{R}_0=1$.

Define $\mathbf{A}(\epsilon)=\mathbf{T}+\mathbf{\Sigma}+\epsilon\mathbf{1}$, where $\mathbf{1}$ is the matrix with all entries equal to one. From similar arguments to those used in the proof of lemma A.3, it follows that the function $\epsilon\mapsto s(\mathbf{A}(\epsilon))$ is monotone increasing. So, if we define $\tilde{\mathbf{A}}(\epsilon)=\mathbf{A}(\epsilon)-s(\mathbf{A}(2\epsilon))\mathbf{I}$, then $s(\tilde{\mathbf{A}}(\epsilon))=s(\mathbf{A}(\epsilon))-s(\mathbf{A}(2\epsilon))\leq 0$. The decomposition $\tilde{\mathbf{A}}(\epsilon)=(\mathbf{T}+\epsilon\mathbf{1})+(\mathbf{\Sigma}-s(\mathbf{A}(2\epsilon))\mathbf{I})$ motivates us to introduce the matrix

$$\mathbf{M}(\epsilon)=-(\mathbf{T}+\epsilon\mathbf{1})(\mathbf{\Sigma}-s(\mathbf{A}(2\epsilon))\mathbf{I})^{-1}.$$

Clearly, $\mathbf{M}(\epsilon)$ converges to $\mathbf{K}_\mathrm{L}$ as $\epsilon\downarrow 0$, and as the spectral radius of $\mathbf{K}_\mathrm{L}$ exceeds one (by lemma A.5), the spectral radius of $\mathbf{M}(\epsilon)$ must be positive for small positive ϵ. Because $\tilde{\mathbf{A}}(\epsilon)$ is clearly irreducible, we use lemma A.4 to deduce that $\rho(\mathbf{M}(\epsilon))\leqslant 1$. ■

Proof of theorem A.1. Combining lemmas A.6 and A.2 (with $\mathcal{R}_0=1$) we conclude that

$$s(\mathbf{T}+\mathbf{\Sigma})=0\quad\Leftrightarrow\quad\mathcal{R}_0=1.$$

By lemma A.4 we have that, at least for small $\epsilon>0$,

$$s(\mathbf{T}+\epsilon\mathbf{1}+\mathbf{\Sigma})<0\quad\Leftrightarrow\quad\rho(-(\mathbf{T}+\epsilon\mathbf{1})\mathbf{\Sigma}^{-1})<1$$

and so, by considering the limit $\epsilon\downarrow 0$ that $s(\mathbf{T}+\mathbf{\Sigma})<0\Rightarrow\mathcal{R}_0\leq 1$ and $\mathcal{R}_0<1\Rightarrow s(\mathbf{T}+\mathbf{\Sigma})\leq 0$. Since, as already noted above, $s(\mathbf{T}+\mathbf{\Sigma})=0\Leftrightarrow\mathcal{R}_0=1$, we conclude that $s(\mathbf{T}+\mathbf{\Sigma})<0\Leftrightarrow\mathcal{R}_0<1$. It follows that $s(\mathbf{T}+\mathbf{\Sigma})>0\Leftrightarrow\mathcal{R}_0>1$, and the proof is complete. ■

ENDNOTES

[1]The word 'typical' is there to emphasize the subtlety that the word 'average' needs to be interpreted in the right way; see Diekmann & Heesterbeek (2000).

[2]As an example consider the standard *SEIR* model. There are two states for infected individuals, the latency state E and the infectious state I. Only the E-state is a state-at-infection, however, because all newly infected individuals start their 'infected life' in state E. One cannot be in the I state immediately after becoming infected, but can only enter state I in the course of the infection. In this example, the NGM only involves the E state, whereas the NGM with large domain involves both infected states.

[3]For completeness, we add that this is a general rule: when an individual can leave a state A, say, in several ways, the probability of going to a particular state B, say, is the product of the *per capita* rate of changing from state A to state B and the average time spent in state A (sojourn time).

[4]In other words: the columns of $\mathbf{E}$ span the range of $\mathbf{T}$.

[5]One way of viewing this property is by saying that there is then only one state-at-infection in a stochastic sense, even though formally there are still two states-at-infection. By 'stochastic sense' we mean that the probability distribution of state-at-infection is fixed, i.e. does not depend on the infectious individual responsible for the transmission.

[6]For completeness we remark that in the decomposition $\mathbf{T}+\mathbf{\Sigma}$ it is essential only that $\mathbf{T}$ is a non-negative matrix and that $\mathbf{\Sigma}$ is a positive off-diagonal matrix with spectral bound $s(\mathbf{\Sigma})<0$ (see appendix A for the terminology). These conditions, however, do not uniquely determine $\mathbf{T}$ and $\mathbf{\Sigma}$. As explained in the text, it is the interpretation that leads to the relevant $\mathbf{T}$ and $\mathbf{\Sigma}$. The interpretation decides which events (production and changes of state) are accounted for in $\mathbf{T}$ and which events in $\mathbf{\Sigma}$. For a concrete example, we refer to Inaba & Nishiura (2008) where in particular the transition from an asymptomatically infected individual to a symptomatically infected individual is considered (as this corresponds more closely to what one can observe). For any decomposition one obtains a 'reproduction number', counting the events incorporated in $\mathbf{T}$. Different decompositions, however, lead to different reproduction numbers. The crucial property explained in appendix A holds for all of them when the conditions above are satisfied.

[7]More generally, one can base the reduction on so-called renewal points in the life cycle; i.e. a subset of states that any individual who will ever reproduce will necessarily visit, and restrict $-(\mathbf{T}\boldsymbol{\Sigma}^{-1})_{ij}$ to that subset of indices. We are, however, not aware of any epidemiologically relevant examples in which the renewal points are neither states-at-infection nor states-of-infectiousness.

[8]Alternatively find the rank of $\mathbf{T}$. This is equal to the number of linearly independent vectors that span the columns of $\mathbf{T}$, so is less than or equal to the number of states-at-infection. If the rank of $\mathbf{T}$ is less than the number of states-at-infection, then $\det \mathbf{K} = 0$. As explained in the text, proceeding in this manner may avoid having to explicitly calculate $\mathbf{K}$.

[9]The function sign is defined in the usual way: $\mathrm{sign}\,(y) = y/|y|$ if $y \neq 0$, and $\mathrm{sign}(0) = 0$.

REFERENCES

Cherry, B. R., Reeves, M. J. & Smith, G. 1998 Evaluation of bovine viral diarrhea virus control using a mathematical model of infection dynamics. *Prev. Vet. Med.* **33**, 91–108. (doi:10.1016/S0167-5877(97)00050-0)

Diekmann, O. & Heesterbeek, J. A. P. 2000 *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation.* Chichester, UK: Wiley.

Diekmann, O., Heesterbeek, J. A. P. & Metz, J. A. J. 1990 On the definition and computation of the basic reproduction ratio $\mathcal{R}_0$ in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365–382. (doi:10.1007/BF00178324)

Inaba, H. & Nishiura, H. 2008 The state-reproduction number for a multistate class age structured epidemic system and its application to the asymptomatic transmission model. *Math. Biosci.* **216**, 77–89. (doi:10.1016/j.mbs.2008.08.005)

Li, C.-K. & Schneider, H. 2002 Applications of Perron–Frobenius theory to population dynamics. *J. Math. Biol.* **44**, 450–462. (doi:10.1007/S002850100132)

Thieme, H. R. 2009 Spectral bound and reproduction number for infinite-dimensional population structure and time heterogeneity. *SIAM J. Appl. Math.* **70**, 188–211. (doi:10.1137/080732870)

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