Bull Math Biol (2011) 73: 777–794
DOI 10.1007/s11538-010-9603-7

ORIGINAL ARTICLE

A Symbolic Investigation of Superspreaders

Chris McCaig · Mike Begon · Rachel Norman · Carron Shankland

Received: 13 May 2009 / Accepted: 2 November 2010 / Published online: 23 December 2010

Abstract Superspreaders are an important phenomenon in the spread of infectious disease, accounting for a higher than average number of new infections in the population. We use mathematical models to compare the impact of supershedders and supercontacters on population dynamics. The stochastic, individual based models are investigated by conversion to deterministic, population level Mean Field Equations, using process algebra. The mean emergent population dynamics of the models are shown to be equivalent with and without superspreaders; however, simulations confirm expectations of differences in variability, having implications for individual epidemics.

Keywords Changing scale · Mean field equation derivation from process algebra

1 Introduction

Traditional models of an epidemic consist of ordinary differential equations (ODEs) that capture the mean change in the number of infected individuals in the population over time (Kermack and McKendrick 1927; Anderson and May 1979). Such models

C. McCaig · R. Norman · C. Shankland (✉)
Department of Computing Science and Mathematics, University of Stirling, Stirling, FK9 4LA, UK
e-mail: ces@cs.stir.ac.uk

C. McCaig
e-mail: cmc@cs.stir.ac.uk

R. Norman
e-mail: ran@cs.stir.ac.uk

M. Begon
School of Biological Sciences, University of Liverpool, Liverpool, L69 3BX, UK
e-mail: mbegon@liverpool.ac.uk

have a well established history, with an associated set of analytical tools. In these models *implicit* assumptions are made about how individual behaviour affects the population as a whole. An alternative method of modelling an epidemic comes in the form of individual based models (Lloyd-Smith et al. 2003). These models are typically studied by stochastic simulation of a population, or by considering the model as a Markov chain problem. Again, there is an associated range of analytical tools. In such models, *explicit* assumptions are made about the behaviour of individuals, which can be based on field observations of a population.

Working at either level of abstraction, individual based or population based, brings the advantages mentioned above; however, if we independently develop *both* types of model, we are left with the problem of how to formally relate the behaviour of individuals to that of the system as a whole. This is important when we wish to know how population level behaviour emerges from individual behaviour. Our previous work (McCaig et al. 2008a, 2009) has developed a method which allows us to write one model (from an individual based perspective) and to automatically derive the other model (the population level). This approach gives the advantages of having both kinds of model, with relationship between the models being formal and explicit. In this particular case of epidemiological modelling, the approach recognises that the transmission process, which is at the heart of infection dynamics, reflects the behaviour of individual hosts, but that for many practical purposes, it is necessary to understand, ideally analytically, the dynamics of the whole system. Conventionally, transmission terms in population models have been derived informally from implicit models of individual behaviour. For example, many models assume that hosts in a population make contact with one another at random, and hence use a simple 'density-dependent' transmission term to reflect a simple linear increase in the rate of contacts with host density. However, while this may suffice for very simple assumptions, such as random mixing, for more complex and biologically realistic patterns of behaviour, transmission terms cannot simply be deduced. Our approach, therefore, allows us to expand the range of biologically realistic transmission behaviours that can be incorporated into population level models, and in this paper, we demonstrate our approach through application to the idea of superspreaders.

Superspreaders are infectious individuals who are somehow responsible for more infections in the population than average (Kemper 1980; Galvani and May 2005; Lloyd-Smith et al. 2005; Matthews and Woolhouse 2005; Fujie and Odagaki 2007; Woolhouse et al. 1997). The 80:20 rule is often cited in this regard, i.e. 20% of the infected individuals are responsible for 80% of further infections. The archetypal superspreader is *Typhoid Mary*. Mary Mallon was a cook in America in the early 1900s (Gibbins 1998). She was exposed to typhoid and became an asymptomatic carrier of the bacteria. Health officials identified her as the source of many typhoid infections, and eventually quarantined her to stop the spread of the disease. She remained incarcerated until her death in 1938. Typhoid Mary is not an unusual case. In July 2008, UK media reported the story of 43 typhoid carriers who had been locked up for life between 1907 and 1992 in an asylum. They were deemed a public health risk (Booth 2008).

Superspreading is also associated with other diseases, including measles and SARS (Lloyd-Smith et al. 2005). Two main hypotheses have been presented regarding the mechanism of superspreading: we will refer to these here as *supershedders*

and *supercontacters*. Supershedders transmit more disease per contact, making subsequent infection more likely, while supercontacters transmit more disease by making more contacts in the population. Two obvious questions arise:

1. Does having superspreaders in a population affect the overall epidemiological dynamics, in particular the form of the transmission term?
2. Does it matter to the formulation of the transmission term what *type* of superspreaders are in the population?

Models of epidemics featuring superspreaders have been addressed to some extent by, for instance, Kemper (1980), Lloyd-Smith et al. (2005), Fujie and Odagaki (2007), but here we are rigorously deriving the population level behaviour from individual interactions. Using individual based modelling, we can express the distinctive behaviour of supershedders and supercontacters. The models can be compared using simulation but we also convert both models to population level Mean Field Equations (MFEs), allowing a more analytical approach. This approach permits investigation of the effect of individual interactions in the individual-based model (expressed using the process algebra Weighted Synchronous Calculus of Communicating Systems (WSCCS)) on emergent population behaviour in the deterministic population-based models (expressed as MFEs). In particular, we can directly investigate the link between the individual interactions of superspreaders and the resultant transmission term.

Background information is given in Sect. 2: Sect. 2.1 presents the notation used for the models, and Sect. 2.2 presents an overview of the technique of McCaig (2007, 2008a) for deriving MFE. Appendices A and B give some additional detail. The core of the paper comprises two models of the different types of superspreaders, and comparison of those models via derived MFE. This is presented in Sect. 3. In order to answer the superspreaders questions above, results from the models are presented in Sect. 3.3. A discussion of those results, and directions for future work, are presented in Sect. 4.

2 Background

Process algebra (Baeten 2005) is one of a range of Computer Science techniques being applied to biological systems. While mathematical models have been used in biology for some time, the computational approach is relatively new, with the majority of applications being in the last ten years. Computer Science techniques can be used to formally express theories about the components of a biological system and the way those components interact. More importantly, just as with mathematical models, those theories can then be explored through computational and analytical methods.

Process algebra has been strongly adopted for use in Systems Biology, e.g. (Priami 2006; Calder and Hillston 2009; Bernardo et al. 2008). Our group has pioneered the use of process algebra for epidemiology (Norman and Shankland 2003; McCaig 2007; McCaig et al. 2009). Process algebras are well suited to describing biological systems which may typically be viewed as networks of (many) interacting components, where the components themselves may have complex, non-deterministic, individual behaviour. In this way, process algebras are similar to Petri

nets. See, e.g. Murata (1989) for an overview. Both approaches have a formal mathematical basis, the advantage of executability, and substantial supporting analytical theory. Petri nets are appealing to use because of their graphical nature, and are particularly useful when true concurrency is required (i.e. actions must occur simultaneously, rather than interleaving concurrency where actions occur discretely, but in any order). For the work presented here, the main advantage of process algebra over Petri nets is an analysis technique based on extraction of Mean Field Equation semantics from process algebra. Process algebra also offers compositional reasoning over models, although this is not utilised here.

2.1 WSCCS Syntax and Semantics

In WSCCS (Weighted Synchronous Calculus of Communicating Systems), the basic components are *actions* and the *processes* (or *agents*) that carry out those actions. The actions are chosen by the modeller to represent activities in the system. For example, *infect*, *send*, *receive*, *throw dice*, and so on. The special pre-defined action $\surd$ simply indicates the passing of time. Processes are constructed via a small number of operators, allowing ordering of actions, probabilistic choice between actions, and parallel composition of processes. The formal syntax and semantics of WSCCS is presented in Tofts (1994), a portion of which is repeated in Appendix A here for easy reference. In Fig. 1, a simple model of an SIR epidemic (Kermack and McKendrick 1927) is presented to illustrate the language.

The model defines twelve agents. The susceptible individuals are modelled by the agents $S1$, $S2$, $S3$, and $SI3$. The infected individuals are modelled by the agents $I1$,

$$
\begin{aligned}
S1 &\stackrel{\text{def}}{=} 1.\surd : S2 \\
I1 &\stackrel{\text{def}}{=} (1-p_{ci}).\surd : I2 + p_{ci}.\surd : I2 \times T2 \\
R1 &\stackrel{\text{def}}{=} 1.\surd : R2 \\
\\
S2 &\stackrel{\text{def}}{=} \omega.\mathit{infect} : SI3 + 1.\surd : S3 \\
I2 &\stackrel{\text{def}}{=} \omega.\mathit{infect} : I3 + 1.\surd : I3 \\
T2 &\stackrel{\text{def}}{=} \omega.\overline{\mathit{infect}} : 0 + 1.\surd : 0 \\
R2 &\stackrel{\text{def}}{=} \omega.\mathit{infect} : R3 + 1.\surd : R3 \\
\\
S3 &\stackrel{\text{def}}{=} 1.\surd : S1 \\
SI3 &\stackrel{\text{def}}{=} p_i.\surd : I1 + (1-p_i).\surd : S1 \\
I3 &\stackrel{\text{def}}{=} p_r.\surd : R1 + (1-p_r).\surd : I1 \\
R3 &\stackrel{\text{def}}{=} 1.\surd : R1 \\
\\
\mathit{Popn} &\stackrel{\text{def}}{=} S1\{s\} \times I1\{i\} \times R1\{r\} \lceil \{\surd\}
\end{aligned}
$$

Fig. 1 Naive SIR model

$I2$, $T2$, and $I3$. The removed individuals are modelled by the agents $R1$, $R2$, and $R3$. The system as a whole (described by *Popn*) comprises s susceptible individuals, i infected individuals, and r removed individuals acting in parallel (the $\times$ operator).

This is a three stage model reflecting three components of infection transmission. In the first stage, the infected individuals have a probabilistic choice to make themselves available for contact or not. In the second stage, contact between individuals happens. In the third stage, contacted susceptibles have a probabilistic choice regarding whether the infection takes hold or not. This reflects three components of disease transmission: probability that a contact between two individuals happens, probability that contact is between a susceptible individual and an infected individual, and probability of getting the disease following such a contact. Note that choices are made probabilistically, and that the agents have no decision making capabilities.

The process which can perform the action a and then evolve to process P is written $a : P$ where a is an action, and P a process. For example, the $S1$ process performs a $\surd$ action and then becomes $S2$. Weighted (probabilistic) choice is expressed with the $+$ operator. For example, process $I3$ can recover with probability p_r (and become the process $R1$) or can continue to be infected with probability $1 - p_r$ (and become the process $I1$). The agent 0 does nothing.

Communication occurs via the paired actions *infect* and $\overline{infect}$. These can be thought of as input and output, respectively (so $T2$ outputs some infection, and $S2$, $I2$, or $R2$ may absorb that infection, with differing results). The special weight ω prioritises communication; if the *infect* action *can* happen, it *must*. WSCCS is a synchronous calculus: in every time step every agent has to perform some action (hence the $\surd$ actions above—these processes are just marking time until the next stage). By combining simple known individuals in parallel in *Popn*, complex overall population level behaviour emerges. The semantics of WSCCS is transition based, yielding a Markov chain interpretation of the model.

A number of analyses are available:

1. stochastic simulation,
2. Markov analysis of the underlying semantics,
3. verification of logical properties,
4. algebraic manipulation of the model.

The first three of these could be computationally expensive, requiring generation of a large underlying state space. Instead, we prefer algebraic manipulation of the model; in this case to transform the model into an equivalent population based model in the form of MFEs.

2.2 Deriving Mean Field Equations

The authors have previously presented (McCaig 2007, 2008a) a method to transform a WSCCS model to MFEs. We do not repeat the method here, but give an overview of benefits and an illustration by application to the simple SIR model of Fig. 1. The method gives an alternative semantics for WSCCS in terms of Mean Field Equations. Algebraic rules are applied to the WSCCS syntax of the model to obtain a set of first-order difference equations expressing the *average* behaviour of the model. Making use of a central limit theorem, first presented by Kurtz (1970), McCaig (2007)

```
function calculateTerm (A, w, a): String {
  case A in {
    probabilistic(A): return w * A_t;
    communicating(A) and priority(A):
      term = (A_t * collaborators(A))/(A_t + competitors(A));
      if a equals √ return (A-term) else return term;
    communicating(A) and not priority(A):
      term = (A_t * collaborators(A))/(A_t + collaborators(A) + competitors(A));
      if a equals √ return (A-term) else return term;
  }}
```

Fig. 2 Pseudo code to calculate proportion of agents at time $t+1$

showed that this approximation to the original transition based semantics offers a close match for large populations and an exact match at the limit, where the overall population size is infinite. There are four benefits to this approach. A new viewpoint of the system is produced, rigorously and symbolically. The resulting MFEs may be amenable to further algebraic analysis using standard mathematical techniques. The problem of handling exponentially increasing state space is avoided. Finally, and to the biologist most importantly, it is possible to exploit known (measured) information about individual behaviour and to link this with emergent population dynamics.

The method is based on algebraic transformation of the syntax of the model. A table is constructed noting the change in the number of each type of agent in the system using the function in Fig. 2. This is a simplified version of term derivation originally presented in McCaig et al. (2008b). Some auxiliary definitions are required. Processes can be classified by syntactic features as: *communicating* (having an action enabled that is involved in a communication), *probabilistic* (having only actions enabled that are not involved in communication), and *priority* (communicating and using ω weights). For a process communicating on action a, we define two groups of agents involved in the synchronisation: *collaborators* are those processes with the matching action $\overline{a}$, and *competitors* are those processes with the same action a. We illustrate the use of the method via a simple example.

Derivation of MFEs for a Simple SIR Model Consider again the simplistic model of disease spread given in Fig. 1. Transition tables track the evolution of numbers of agents, and are indexed by $(agent1, action) \times (agent2)$. An entry indicates the number of $agent1$ evolving to $agent2$, by performing the action. For example, all $S1_t$ (the $S1$ agents at time t) evolve to $S2_{t+1}$ (the $S2$ agents at time $t+1$), but only $p_{ci} I1_t$ of $I1_t$ evolve to $T2_{t+1}$.

The populated parts of the transition table for the system of Fig. 1 are in Tables 1, 2, 3. Each column leads to a MFE for that agent, but 0 is ignored here since this is not of interest to us. The method outlined above generates the following MFEs:

$$\begin{aligned} rS_{t+1} &= S_t - \frac{p_i S_t I_t}{N_t}, \\ I_{t+1} &= (1-p_r) I_t + \frac{p_i S_t I_t}{N_t}, \\ R_{t+1} &= R_t + p_r I_t. \end{aligned} \tag{1}$$

Table 1 Transition table for SIR model of Fig. 1 (first stage transitions)

	$S2_{t+1}$	$I2_{t+1}$	$T2_{t+1}$	$R2_{t+1}$
$(S1_t, \surd)$	$S1_t$			
$(I1_t, \surd)$		$(1-p_{ci})I1_t + p_{ci}I1_t$	$p_{ci}I1_t$	
$(R1_t, \surd)$				$R1_t$

Table 2 Transition table for SIR model of Fig. 1 (second stage transitions)

	0	$S3_{t+1}$	$SI3_{t+1}$	$I3_{t+1}$	$R3_{t+1}$
$(S2_t, infect)$			$\frac{S2_t T2_t}{(S2_t+I2_t+R2_t)}$		
$(S2_t, \surd)$		$S2_t - \frac{S2_t T2_t}{(S2_t+I2_t+R2_t)}$			
$(I2_t, *)$				$I2_t$	
$(T2_t, *)$	$T2_t$				
$(R2_t, *)$					$R2_t$

Table 3 Transition table for SIR model of Fig. 1 (third stage transitions)

	$S1_{t+1}$	$I1_{t+1}$	$R1_{t+1}$
$(S3_t, \surd)$	$S3_t$		
$(SI3_t, \surd)$	$(1-p_i)SI3_t$	$p_i SI3_t$	
$(I3_t, \surd)$		$(1-p_r)I3_t$	$p_r I3_t$
$(R3_t, \surd)$			$R3_t$

Equations for $S1_{t+3}, I1_{t+3}, R1_{t+3}$ in terms of $S1_t, I1_t, R1_t$ are produced by substitution. These are rewritten as one stage difference equations to give (1), since we are not interested in the intermediate stages of the model.

3 Models

The models presented below are variations on the basic SIR model given in Fig. 1, with the addition of births and deaths (for biological realism), and of course, superspreaders. In both cases, the superspreaders are added as a new type of infected individual U which has different behaviour to the existing infected individual. Death due to the disease is ignored, but is easily added if required.

3.1 Supercontacters

In the model of Fig. 3, the superspreader is a *supercontacter*. That is, this individual is more gregarious and makes more contacts with the rest of the population than the average infected individual. This is modelled here by setting a special contact rate for supercontacters: $p_{cu} = \alpha p_{ci}$, where $\alpha \in \mathbb{R}$ is the supercontacter multiplier, $\alpha > 1$. In other words, supercontacters are more likely to make contact.

$$
\begin{aligned}
p_b &\overset{\text{prob}}{=} p_{b_0} - k(\lfloor S1 \rfloor + \lfloor I1 \rfloor + \lfloor U1 \rfloor + \lfloor R1 \rfloor) \\
\\
S1 &\overset{\text{def}}{=} p_b.\surd : S2 \times B2 + (1 - p_b).\surd : S2 \\
I1 &\overset{\text{def}}{=} p_b(1 - p_{ci}).\surd : I2 \times B2 + p_b p_{ci}.\surd : I2 \times T2 \times B2 \\
&\quad + (1 - p_b)(1 - p_{ci}).\surd : I2 + (1 - p_b) p_{ci}.\surd : I2 \times T2 \\
U1 &\overset{\text{def}}{=} p_b(1 - p_{cu}).\surd : U2 \times B2 + p_b p_{cu}.\surd : U2 \times T2 \times B2 \\
&\quad + (1 - p_b)(1 - p_{cu}).\surd : U2 + (1 - p_b) p_{cu}.\surd : U2 \times T2 \\
R1 &\overset{\text{def}}{=} p_b.\surd : R2 \times B2 + (1 - p_b).\surd : R2 \\
\\
S2 &\overset{\text{def}}{=} \omega.\mathit{infect} : SI3 + 1.\surd : S3 \\
I2 &\overset{\text{def}}{=} \omega.\mathit{infect} : I3 + 1.\surd : I3 \\
T2 &\overset{\text{def}}{=} \omega.\overline{\mathit{infect}} : 0 + 1.\surd : 0 \\
U2 &\overset{\text{def}}{=} \omega.\mathit{infect} : U3 + 1.\surd : U3 \\
R2 &\overset{\text{def}}{=} \omega.\mathit{infect} : R3 + 1.\surd : R3 \\
B2 &\overset{\text{def}}{=} 1.\surd : B3 \\
\\
S3 &\overset{\text{def}}{=} (1 - p_d).\surd : S1 + p_d.\surd : 0 \\
SI3 &\overset{\text{def}}{=} p_i(1 - p_s).\surd : I1 + p_i p_s.\surd : U1 \\
&\quad + (1 - p_d - p_i).\surd : S1 + p_d.\surd : 0 \\
I3 &\overset{\text{def}}{=} p_r.\surd : R1 + (1 - p_r - p_d).\surd : I1 + p_d.\surd : 0 \\
U3 &\overset{\text{def}}{=} p_r.\surd : R1 + (1 - p_r - p_d).\surd : U1 + p_d.\surd : 0 \\
R3 &\overset{\text{def}}{=} (1 - p_d).\surd : R1 + p_d.\surd : 0 \\
B3 &\overset{\text{def}}{=} 1.\surd : S1 \\
\\
\mathit{Popn\_C} &\overset{\text{def}}{=} (S1\{s\} \times I1\{i\} \times U1\{u\} \times R1\{r\}) \lceil \{\surd\}
\end{aligned}
$$

Fig. 3 Contact superspreader model. I and U make at most 1 contact per iteration with probabilities p_{ci} and $p_{cu} = \alpha p_{ci}$, respectively

This is not the only way to express that an individual makes more contacts (McCaig 2007). For example, the supercontacter may have the same p_{ci} as the infected but evolve to $TU2$ in which multiple *infect* actions can be performed. That model produces the same results as here, with the constraint that the multiplying factor must be integer (the number of actions can only be a positive integer). Here, α can be non-integer.

Stage 1 ($S1$, $I1$, $U1$, $R1$) is a birth stage. All agents reproduce with probability p_b. Birth is density dependent (as described in McCaig et al. 2008b). Newborns are not available for infection in subsequent stages ($B2$, $B3$). Additionally, the infected and supercontacter agents probabilistically become available to contact others (so only a subset of infected individuals try to make new infections in the next stage), with

probabilities p_{ci} and $p_{cu,}$ respectively. Stage 2 is the infection stage. Communication happens between infected individuals of either type and the rest of the population.

In stage 3, the agents $SI3$, which have come into contact with the infection, become infected with probability p_i. A probabilistic choice is also made as to whether the new infected individual is a supercontacter or not (with probability p_s). Lastly, agents may die of natural causes (with probability p_d) or recover from illness (with probability p_r).

The MFEs arising from the model in Fig. 3 are

$$\begin{aligned} S_{t+1} &= (1-p_d)S_t - \frac{p_i p_{ci} S_t (I_t + \alpha U_t)}{N_t} + (p_{b_0} - kN_t)N_t, \\ I_{t+1} &= (1-p_d-p_r)I_t + \frac{p_i(1-p_s)p_{ci} S_t (I_t + \alpha U_t)}{N_t}, \\ U_{t+1} &= (1-p_d-p_r)U_t + \frac{p_i p_s p_{ci} S_t (I_t + \alpha U_t)}{N_t}, \\ R_{t+1} &= (1-p_d)R_t + p_r(I_t + U_t). \end{aligned} \tag{2}$$

3.2 Supershedders

In the model of Fig. 4, the superspreader is a *supershedder*. That is, following infection, this individual delivers more infection to the rest of the population per contact. Some authors have hypothesised that this is due to genetic factors influencing, for example the shape of the throat. Another hypothesis is that these individuals have an altered or compromised immune system, either intrinsically (genetic differences between individuals) or perhaps as a result of co-infection with another pathogen (HIV-AIDS being a notable example in humans). Supershedding is modelled here by setting a special infection rate for supershedders: $p_{iu} = \alpha p_i$, where $\alpha \in \mathbb{R}$ is the supershedder multiplier.

The model is constructed in three stages in much the same way as the model of Fig. 3. In this case, both types of infecteds are equally likely to make an infectious contact, but as mentioned in Sect. 2.2 a different communication action is used for supershedders, to allow differentiation between contact with a supershedder and contact with a normal infected individual. This is important because in stage 3 agents $SU3$ have been contacted by a supershedder and get the infection with probability p_{iu}. Agents $SI3$ have been contacted by a normal infected individual and get the infection with probability p_i. An alternative modelling approach is to allow infected individuals and supershedding individuals to try to infect in separate steps, but the solution presented here is felt to be more intuitive in terms of expressing individual behaviour.

Technically, the method of McCaig et al. (2008a) does not apply here because two different actions occur in the same step. Appendix B details the extension to the method required. The model in Fig. 4 leads to (2), the same MFEs as for Fig. 3.

3.3 Results

Two models of the superspreading phenomenon have been presented. The models are rather different in individual behaviour and some difference in population dynamics

$$
\begin{aligned}
p_b &\overset{\text{prob}}{=} p_b 0 - k * (\lfloor S1 \rfloor + \lfloor I1 \rfloor + \lfloor U1 \rfloor + \lfloor R1 \rfloor) \\
\\
S1 &\overset{\text{def}}{=} p_b.\surd : S2 \times B2 + (1 - p_b).\surd : S2 \\
I1 &\overset{\text{def}}{=} p_b(1 - p_{ci}).\surd : I2 \times B2 + p_b p_{ci}.\surd : I2 \times T2 \times B2 \\
&\quad + (1 - p_b)(1 - p_{ci}).\surd : I2 + (1 - p_b) p_{ci}.\surd : I2 \times T2 \\
U1 &\overset{\text{def}}{=} p_b(1 - p_{ci}).\surd : U2 \times B2 + p_b p_{ci}.\surd : U2 \times TU2 \times B2 \\
&\quad + (1 - p_b)(1 - p_{ci}).\surd : U2 + (1 - p_b) p_{ci}.\surd : U2 \times T2 \\
R1 &\overset{\text{def}}{=} p_b.\surd : R2 \times B2 + (1 - p_b).\surd : R2 \\
\\
S2 &\overset{\text{def}}{=} \omega.infect : SI3 + \omega.infectU : SU3 + 1.\surd : S3 \\
I2 &\overset{\text{def}}{=} \omega.infect : I3 + \omega.infectU : I3 + 1.\surd : I3 \\
T2 &\overset{\text{def}}{=} \omega.\overline{infect} : 0 + 1.\surd : 0 \\
U2 &\overset{\text{def}}{=} \omega.infect : U3 + \omega.infectU : U3 + 1.\surd : U3 \\
TU2 &\overset{\text{def}}{=} \omega.\overline{infectU} : 0 + 1.\surd : 0 \\
R2 &\overset{\text{def}}{=} \omega.infect : R3 + \omega.infectU : R3 + 1.\surd : R3 \\
B2 &\overset{\text{def}}{=} 1.\surd : B3 \\
\\
S3 &\overset{\text{def}}{=} (1 - p_d).\surd : S1 + p_d.\surd : 0 \\
SI3 &\overset{\text{def}}{=} p_i(1 - p_s).\surd : I1 + p_i p_s.\surd : U1 \\
&\quad + (1 - p_d - p_i).\surd : S1 + p_d.\surd : 0 \\
SU3 &\overset{\text{def}}{=} p_{iu}(1 - p_s).\surd : I1 + p_{iu} p_s.\surd : U1 \\
&\quad + (1 - p_d - p_{iu}).\surd : S1 + p_d.\surd : 0 \\
I3 &\overset{\text{def}}{=} p_r.\surd : R1 + (1 - p_d - p_r).\surd : I1 + p_d.\surd : 0 \\
U3 &\overset{\text{def}}{=} p_r.\surd : R1 + (1 - p_d - p_r).\surd : U1 + p_d.\surd : 0 \\
R3 &\overset{\text{def}}{=} (1 - p_d).\surd : R1 + p_d.\surd : 0 \\
B3 &\overset{\text{def}}{=} 1.\surd : S1 \\
\\
Popn\_S &\overset{\text{def}}{=} (S1\{s\} \times I1\{i\} \times U1\{u\} \times R1\{r\}) \lceil \{\surd\}
\end{aligned}
$$

Fig. 4 Supershedder model with density dependent probability of giving birth

may be expected; however, the derived mean field equations (2) are identical. As expected simulation of the models also gives the same mean behaviour at the population level. In Fig. 5, we plot the mean of 1,000 simulations of each model. The parameters here were chosen to reflect the 80:20 rule of infection by superspreaders. Equivalence of both the MFEs and the simulations suggests that the particular *mechanisms* for superspreading are not important if we are only interested in average behaviour of the population.

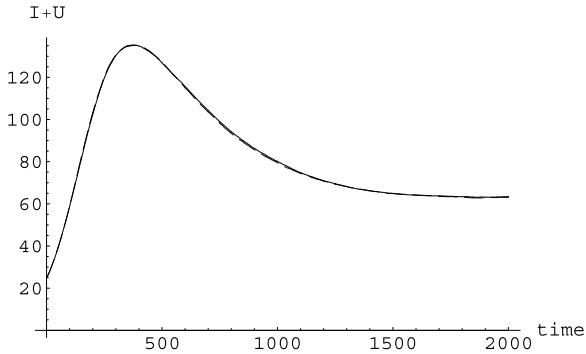

Fig. 5 Mean of total infecteds $(I+U)$ of 1,000 simulations with $p_{b0}=0.2$, $k=0.0008$, $p_{ci}=0.06$, $p_d=0.001$, $p_i=0.06$, $\alpha=16$, $p_s=0.2$, $p_r=0.002$, $S_0=100$, $I_0=20$, $U_0=5$: _ _ _ Fig. 3, ____ Fig. 4

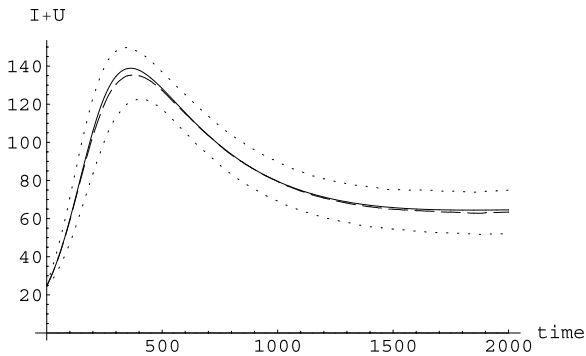

Fig. 6 Total infecteds $(I+U)$ for Fig. 4 with $p_{b0}=0.2$, $k=0.0008$, $p_{ci}=0.06$, $p_d=0.001$, $p_i=0.06$, $\alpha=16$, $p_s=0.2$, $p_r=0.002$, $S_0=100$, $I_0=20$, $U_0=5$: ____ MFE; Simulations _ _ _ mean, . . . mean $\pm$ SD

It has been shown (McCaig 2007; McCaig et al. 2008a) that in the limiting case, where the total population size is infinite, MFEs will exactly match the mean behaviour of a model. Figure 6 plots the mean of 1,000 simulations of the model of Fig. 4 and the time series of the MFEs. The MFEs offer a very good approximation to the mean of the simulations, lying well within the region defined by one standard deviation either side of the mean of the simulations. An almost identical graph would be produced for the model of Fig. 3 (given Fig. 5).

In addition to the MFEs and the simulations being equivalent, further endorsement of our result may be obtained through the literature. Kemper (1980) proposed an ODE model of a system featuring superspreaders. By removing the terms for birth and death from (2) our MFEs match Kemper's ODEs under the following transformation of our parameters: $U=I_1$, $I=I_2$, $p_i p_{ci}\alpha/N=r_1$, $p_i p_{ci}/N=r_2$, $p_s=\beta$, $p_r=\gamma$.

It is reasonable to ask how the models of Figs. 3 and 4 compare with a similar SIR model *without* superspreaders. The MFEs (2) can be shown equivalent to those of the SIR model, with modified parameters p_i', p_{ci}' and α'. This is done by setting $\alpha'=1$, so that I and U have the same behaviour, and equating transmission in the non-superspreader SIR model with transmission in the superspreader model, such that

$$\frac{p_i' p_{ci}' S_t(I_t+U_t)}{N_t}=\frac{p_i p_{ci} S_t(I_t+\alpha U_t)}{N_t}. \tag{3}$$

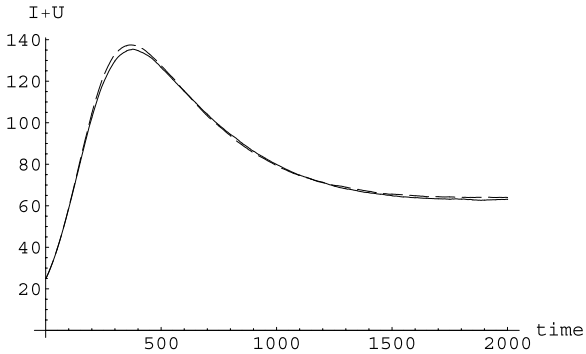

Fig. 7 Mean for models with and without superspreaders with $p_{b0} = 0.2$, $k = 0.0008$, $p_{ci} = 0.06$ ($p'_{ci} = 0.24$), $p_d = 0.001$, $p_i = 0.06$ ($p'_i = 0.06$), $\alpha = 16$ ($\alpha' = 1$), $p_s = 0.2$, $p_r = 0.002$, $S_0 = 100$, $I_0 = 20$, $U_0 = 5$: Simulations ____ without superspreaders, _ _ _ with superspreaders

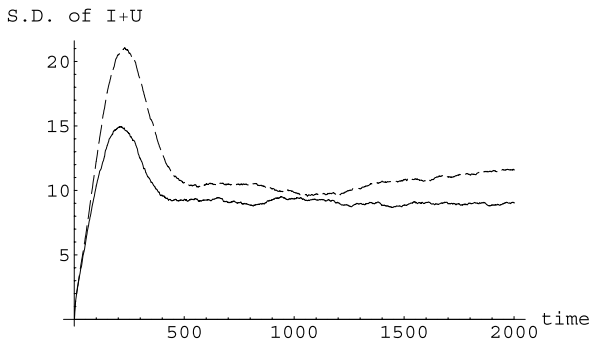

Fig. 8 Standard deviation for models with and without superspreaders with $p_{b0} = 0.2$, $k = 0.0008$, $p_{ci} = 0.06$ ($p'_{ci} = 0.24$), $p_d = 0.001$, $p_i = 0.06$ ($p'_i = 0.06$), $\alpha = 16$ ($\alpha' = 1$), $p_s = 0.2$, $p_r = 0.002$, $S_0 = 100$, $I_0 = 20$, $U_0 = 5$: Simulations ____ without superspreaders, _ _ _ with superspreaders

By noting from (2) that $U_t = p_s(I_t + U_t)$ and $I_t = (1 - p_s)(I_t + U_t)$ we can simplify (3):

$$\frac{p'_i p'_{ci} S_t (I_t + U_t)}{N_t} = \frac{p_i p_{ci} S_t((1 - p_s)(I_t + U_t) + \alpha p_s (I_t + U_t))}{N_t},$$

$$p'_i p'_{ci} = p_i p_{ci}(1 - p_s + \alpha p_s),$$

i.e. the probabilities of making contact and becoming infected after contact are merely rescaled in the non-superspreader model to achieve the same mean behaviour.

At this point, we ask do superspreaders make *any* difference to the models? We expect that by introducing more individual variability that variability at the population level would also increase. This can only be seen through simulation results. In Figs. 7 and 8, we present the results of simulations of the two models (with and without superspreaders). We can see in Fig. 7 that the mean of the simulations for these two models is almost identical; however, in Fig. 8 the standard deviations are different. The peak of the standard deviations in both cases, and the biggest difference between the two, is at around $t = 200$, which corresponds to the period when the number infected is rising most rapidly. This increased variability could play an important role in determining whether a given realisation of the epidemic will become very large, or die out before it is established in the population.

Ideally, we would like to produce an approximation for the standard deviation in much the same way as the MFE approximate the mean. Developing such a method would be an important piece of future work to extend our method.

4 Summary, Conclusions and Future Work

We began by asking:

1. Does having superspreaders in a population affect the overall epidemiological dynamics, in particular the form of the transmission term?
2. Does it matter to the formulation of the transmission term what type of superspreaders are in the population?

Through the models presented in this paper, we have shown that the answer to the second question, given the assumptions that we have made, is "no". Despite differences in individual behaviour, the models of supershedders and supercontacters have the same mean behaviour. However, if we think beyond the model to the practical aspects of disease control then it is likely that the differences will be very important. Lloyd Smith et al. (2005) suggest that control efforts should be aimed at identifying the superspreaders in order to control an outbreak more quickly, and for sexually transmitted diseases at least, Cohen et al. (2003) have suggested a mechanism for doing this. However, identifying superspreaders is much easier if they are supercontacters (for example, gregarious individuals, or those who have many sexual partners) than if they are supershedders, unless their supershedding is associated with an identifiable pathology.

Further, the similarities between the transmission rates in the two superspreader models present another problem. The utility of models is usually demonstrated through matching with historical epidemic data, e.g. as we have done for AIDS in McCaig et al. (2009). In this case, both superspreader models would match data equally well. Moreover, the non-superspreader model would provide a similar match. This means that if we are only interested in the mean behaviour of the system, then we could argue that the simpler non-superspreader model would be the most parsimonious and is therefore the one that should be used. In addition, we can see from (3) that if we estimated p_i' and p_{ci}' for the non-superspreader model under circumstances where superspreaders exist then we would overestimate these terms. This could be argued to be an advantage if were to think about control because it would mean that we would overestimate the amount of control needed and therefore would be more likely to control the disease. However, if we *are* able to carry out targeted treatment on superspreaders then that could be much more effective. In this case the best solution may depend on the "strength" of the superspreader and our ability to identify them. As we have said, if we are only interested in mean behaviour then the non-superspreader model will do just as well. However, there is a significant difference between the models when we consider the amount of variability within the stochastic simulations. We can see from Fig. 8 that the superspreader models have more variability within the simulations than the non-superspreader models, especially in the early stages. This is not surprising since in the early stages of an epidemic the dynamics can change radically depending on whether or not one of the first individuals

infected is a superspreader. This was discussed in more detail in Galvani and May (2005). This means that it becomes much more difficult to predict the course of a single epidemic when there are superspreaders present. Therefore, the answer to the first question is "no" if we are only interested in the mean, but for almost all practical purposes the answer is "yes".

Two strands of further work can be carried out, one with a biological emphasis, the other oriented to Theoretical Computer Science. A useful question to ask is: have we captured supershedders and supercontacters adequately in our models? In fact, we have experimented with several different ways of presenting the models. Aside from the modelling choices already mentioned in Sect. 3, it could be regarded, for example, that there should be a class of supercontacting susceptible individuals as well as supercontacting infected individuals (we do not expect their behaviour to change on getting the disease). This quickly leads to a model in which only supercontacting individuals get the disease, so is rejected. Likewise, is there a supersusceptible group (with compromised immune systems perhaps) who might turn into supershedders? As above, this would produce a subclass of the population with the disease, concentrating on the supersusceptibles and supershedders.

We have shown that the two models presented in this paper are equivalent in terms of MFE, yet they are not equivalent under any of the usual process algebra equivalences (since they have different actions, and the branching probabilities are different). An interesting development in terms of Theoretical Computer Science and Mathematics might be to define an equivalence relation for WSCCS based on mean field equation semantics.

The application of theoretical computer science techniques to biological systems is still at an early stage of development. We have shown here that by using process algebra to describe the model in terms of individual behaviour, we can rigorously derive a population level model, allowing investigation of the relationship between individual interactions and transmission dynamics. We see this as a major benefit of using process algebra, but there are others: using a process algebra gives access to a range of ways to explore a model, each lending different insights to overall system behaviour. This ability will become even more useful when investigating more complex systems.

Acknowledgements This work was supported by EPSRC through a Doctoral Training Grant (CM, from 2004–2007), and through *System Dynamics from Individual Interactions: A process algebra approach to epidemiology* (EP/E006280/1, all authors, 2007–2010). We are grateful to the anonymous referees for their helpful comments.

Appendix A: Syntax of WSCCS

The possible WSCCS expressions are given by the following BNF grammar:

$$A ::= X \mid a{:}A \mid \Sigma\{w_i.A_i | i \in I\} \mid A \times B \mid A\lceil L \mid \Theta(A) \mid A[S] \mid X \stackrel{\text{def}}{=} A.$$

Here, $X \in Var$, a set of process variables; $a \in Act$, an action group; $w_i \in \mathcal{W}$, a set of weights; S a set of renaming functions, $S : Act \to Act$ such that $S(\surd) = \surd$ and

$\overline{S(a)} = S(\overline{a})$; action subsets $A \subseteq Act$ with $\surd \in A$; and arbitrary indexing sets I. Actions form an Abelian group with identity $\surd$ and the inverse of action a being $\overline{a}$. Actions occur instantaneously and have no duration.

The informal interpretation of the operators is as follows:

- 0 a process which cannot proceed, representing deadlock;
- X the process bound to the variable X;
- $a{:}A$ a process which can perform the action a becoming the process A;
- $\Sigma\{w_i.A_i | i \in I\}$ the weighted choice between processes A_i, the weight of A_i being w_i. Considering a large number of repeated experiments of this process, we expect to see A_i chosen with relative frequency $w_i / \Sigma_{i \in I} w_i$. Weights are generally positive natural numbers or reals, but may also incorporate the special weight ω which is greater than all natural numbers. This is used in *priority* and is written $m\omega^n$ where $m, n \geq 0$. The binary plus operator can be used in place of the indexed sum, i.e. writing $\Sigma\{1_1.a{:}0, 2_2.b{:}0 | i \in \{1, 2\}\}$ as $1.a{:}0 + 2.b{:}0$;
- $A \times B$ the synchronous parallel composition of A and B. At each stage, each process must perform an action with the composed process performing the composition (denoted #) of the individual actions, e.g. $a{:}A \times b{:}B$ yields $a\#b{:}(A \times B)$. This is a powerful operator: models are constructed by describing simple individuals and composing a number of those in parallel. Here, we use an extended notation (McCaig 2007) $A\{n\}$ which is syntactic sugar for n instances of process A in parallel, where $n \in \mathbb{N}$;
- $A \lceil L$ a process which can only perform actions in the group L. This operator is used to enforce communication on actions $b \notin L$. Two processes in parallel may communicate when one carries out an action and the other carries out the matching co-action, e.g. *infect* and $\overline{infect}$. Communication can be used to model passing of information from one process to another, or to coordinate activity. Such communication is strictly two-way; that is, only two processes may interact on this action;
- $\Theta(A)$ represents taking the prioritised parts of the process A only;
- $A[S]$ represents A relabelled by the function S (we do not use relabelling in this paper, but it is included for completeness);
- $X \stackrel{\text{def}}{=} A$ represents binding the process variable X to the expression A.

Appendix B: Multiple Alternative Communicating Actions

The method of McCaig (2007) applies to a subset of WSCCS models. One of the restrictions imposed is that only a single communicating action may be presented in each communication step. In the model of Fig. 4, two communicating actions are in the same step: this is required to distinguish supershedders and normal infecteds. This section presents a general extension to the method to handle agents such as $S2$, $I2$, $U2$, and $R2$ of Fig. 4.

The general form of the agent is

$$A = \omega.a_1 : A1 + \omega.a_2.A2 + 1.b.A3.$$

Two sets of collaborating agents $C1$ and $C2$ perform the actions $\overline{a_1}$ and $\overline{a_2}$, respectively. Communication is prioritised. The agent A can perform either a_1 or a_2, evolving differently in each case, but cannot perform both actions together. The action b is a non-communicating action and because of priority will only be executed if neither a_1 nor a_2 can synchronise with another process. There may be other processes able to collaborate with $C1$ and $C2$. These are the competitors of A. The total number of agents doing the a action, i.e. the A_t agents plus their competitors, is denoted N_t. The extension to the method calculates the number of A agents communicating with $C1$ agents and becoming $A1$, and the number of A agents communicating with $C2$ agents and becoming $A2$.

In the following, the multi-nomial coefficient $\binom{m}{p,q,r}$ is used. This represents the number of unordered ways to choose a group of p objects, a group of q objects and a group of r objects from a group of m distinct objects, with $m = p + q + r$.

The number of A_t which communicate with $C1_t$ is

$$\frac{\sum_k \sum_j k\binom{A_t}{k,j,A_t-k-j}\binom{N_t-A_t}{C1_t-k,C2_t-j,N_t-A_t-C1_t-C2_t+k+j}}{\sum_k \sum_j \binom{A_t}{k,j,A_t-k-j}\binom{N_t-A_t}{C1_t-k,C2_t-j,N_t-A_t-C1_t-C2_t+k+j}}. \tag{4}$$

On the numerator, we have the weighted sum of all possible evolutions of A agents to $A1$ agents. That is, if the evolution is to a state with 42 A agents, then we multiply the likelihood of getting to that state by 42. Similarly, if the evolution is to a state with a single A agent, then we multiply by 1. This is k in the expression above. The second component of the numerator indicates the number of A communicating with $C1$ agents, and the third component indicates the number of competitors of A communicating with $C1$ agents. The denominator is the same sum, un-weighted, representing all possible evolutions of A.

Fortunately (4) can be simplified using Vandermonde's convolution (Graham et al. 1989), yielding

$$\frac{A_t C1_t}{N_t}.$$

This term is valid only when $N_t \geq C1_t + C2_t$. If $N_t < C1_t + C2_t$ then there are more actions from $C1$ and $C2$ than there are from A and its competitors. In this case, the number of A communicating with $C1$ is

$$\frac{A_t C1_t}{C1_t + C2_t}.$$

Therefore, the general term for the number of A_t agents which communicate with $C1_t$ is

$$calculateTerm(A, w, a_1) = \min\left(\frac{A_t C1_t}{N_t}, \frac{A_t C1_t}{C1_t + C2_t}\right).$$

The result for two actions can be generalised to cover cases where there are n different actions $a_1, a_2, \ldots, a_n$, and rephrased in the language of Fig. 2, giving

$$\begin{aligned} &calculateTerm(A, w, a_m) \\ &\quad = \min\big(\big(A_t * collaborators(A, a_m)\big)/\big(A_t + competitors(A, a_m)\big), \\ &\qquad \big(A_t * collaborators(A, a_m)\big)/\big(sum\big(i, collaborators(A, a_i)\big)\big)\big) \end{aligned}$$

where m ranges over $1 \ldots n$. The auxiliary function $collaborators(A, a_m)$ denotes the set of agents collaborating on action a_m (similarly for competitors), and *sum(i,expression)* iterates over *expression* for all values of i.

References

Anderson, R. M., & May, R. M. (1979). Population biology of infectious-diseases. 1. *Nature*, *280*, 361–367.

Baeten, J. C. M. (2005). A brief history of process algebra. *Theor. Comput. Sci.*, *335*(2/3), 131–146.

Bernardo, M., Degano, P., & Zavattaro, G. (Eds.) (2008). *Lecture notes in computer science: Vol.* 5016. *Formal methods for computational systems biology*. Berlin: Springer.

Booth, J. (2008). Britain's Typhoid Marys locked up for life in an Epsom asylum. *The Times*, 28 July 2008. Available at http://www.timesonline.co.uk/tol/news/uk/health/article4414995.ece (Accessed: 2/2/2010).

Calder, M., & Hillston, J. (2009). Process algebra modelling styles for biomolecular processes. In *Lecture notes in computer science: Vol.* 5750. *Transactions on computational systems biology XI* (pp. 1–25).

Cohen, R., Havlin, S., & ben Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.*, *91*(24), 247901.

Fujie, R., & Odagaki, T. (2007). Effects of superspreaders in spread of epidemic. *Phys. A Stat. Mech. Appl.*, *374*, 843–852.

Galvani, A. P., & May, R. M. (2005). Epidemiology—dimensions of superspreading. *Nature*, *438*(7066), 293–295.

Gibbins, L. N. (1998). Mary Mallon: disease denial, and detention. *J. Biol. Educ.*, *32*, 127–132.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1989). *Concrete mathematics: a foundation for computer science*. Reading: Addison-Wesley.

Kemper, J. T. (1980). Identification of superspreaders for infectious-disease. *Math. Biosci.*, *48*, 111–127.

Kermack, W. O., & McKendrick, A. G. (1927). Contributions to the mathematical theory of epidemics i. *Proc. R. Soc. Lond. Ser. A*, *115*, 700–721.

Kurtz, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.*, *7*, 49–58.

Lloyd-Smith, J. O., Galvani, A. P., & Getz, W. M. (2003). Curtailing transmission of severe acute respiratory syndrome within a community and its hospital. *Proc. R. Soc. Lond. Ser. B*, *270*(1528), 1979–1989.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, *438*, 355–359.

Matthews, L., & Woolhouse, M. (2005). New approaches to quantifying the spread of infection. *Nat. Rev. Microbiol.*, *3*, 529–536.

McCaig, C. (2007). *From individuals to populations: changing scale in process algebra models of biological systems*. Ph.D. thesis, University of Stirling. http://hdl.handle.net/1893/398.

McCaig, C., Norman, R., & Shankland, C. (2008a). *Deriving mean field equations from large process algebra models* (Technical Report CSM-175). Department of Computing Science and Mathematics, University of Stirling, March 2008. http://hdl.handle.net/1893/1584.

McCaig, C., Norman, R., & Shankland, C. (2008b). Process algebra models of population dynamics. In *Lecture notes in computer science: Vol.* 5147. *Algebraic biology* (pp. 139–155). Berlin: Springer.

McCaig, C., Norman, R., & Shankland, C. (2009). From individuals to populations: a symbolic process algebra approach to epidemiology. *Math. Comput. Sci.*, *2*(3), 139–155.

Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proc. IEEE*, *27*(4), 541–580.
Norman, R., & Shankland, C. (2003). Developing the use of process algebra in the derivation and analysis of mathematical models of infectious disease. In *Lecture notes in computer science: Vol.* 2809. *Computer aided systems theory—EUROCAST 2003* (pp. 404–414). Berlin: Springer.
Priami, C. (2006). Process calculi and life science. *Electron. Notes Theor. Comput. Sci.*, *162*, 301–304.
Tofts, C. (1994). Processes with probabilities, priority and time. *Form. Asp. Comput.*, *6*, 536–564.
Woolhouse, M. E. J., Dye, C., Etard, J. F., Smith, T., Charlwood, J. D., Garnett, G. P., Hagan, P., Hii, J. L. K., Ndhlovu, P. D., Quinnell, R. J., Watts, C. H., Chandiwana, S. K., & Anderson, R. M. (1997). Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proc. Natl. Acad. Sci. USA*, *94*, 338–342.