Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 247 (2007) 765–774
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Emergent heterogeneity in declining tuberculosis epidemics

Caroline Colijn[a,*], Ted Cohen[a,b], Megan Murray[a,b,c]

[a]Department of Epidemiology, Harvard School of Public Health, USA
[b]Department of Social Medicine and Health Inequalities, Brigham and Women's Hospital, USA
[c]Infectious Disease Unit, Massachusetts General Hospital, USA

Received 11 December 2006; received in revised form 20 April 2007; accepted 23 April 2007
Available online 27 April 2007

Abstract

Tuberculosis is a disease of global importance: over 2 million deaths are attributed to this infectious disease each year. Even in areas where tuberculosis is in decline, there are sporadic outbreaks which are often attributed either to increased host susceptibility or increased strain transmissibility and virulence. Using two mathematical models, we explore the role of the contact structure of the population, and find that in declining epidemics, localized outbreaks may occur as a result of contact heterogeneity even in the absence of host or strain variability. We discuss the implications of this finding for tuberculosis control in low incidence settings.

Keywords: Epidemiology; Networks; Mathematical model; Delay model; Reinfection

1. Introduction

Tuberculosis (TB) is caused by infection with the bacterium *Mycobacterium tuberculosis*. The bacilli spread through the respiratory route: individuals with active disease may transmit infection if the airborne particles produced when they cough, talk, and sing are inhaled by others. Once infected, individuals enter a period of latency during which they exhibit no symptoms and are not infectious to others. While most are able to contain this infection indefinitely, at least 10% will eventually progress to disease and expose others (Sutherland et al., 1982, 1976; Styblo, 1991). Although approximately one-third of the world's population harbors a latent *M. tuberculosis* infection (Dye, 2006), this statistic belies the great heterogeneity in risk among individuals and among different countries. In some areas the lifetime risk of infection nears 100% while in others the probability of exposure is minimal.

Mathematical modeling has proven a valuable tool for understanding TB dynamics (Blower et al., 1995; Vynnycky and Fine, 1997; Feng et al., 2000; Singer and Kirschner, 2004) and has served as the basis for establishing control targets and assessing policy strategies (Blower et al., 1996; Dye et al., 1998; Cohen et al., 2006). However, most such models, with occasional exceptions (Schinazi, 1999), have been differential equation susceptible-exposed-infected-recovered (SEIR) models that assume a homogenously mixed population. In populations where people contact only a small subset of the population (such as their colleagues, friends, families, etc.), respiratory diseases such as TB are more likely to be transmitted among local groups of contacts. Non-random mixing introduces "contact structure", which is defined here as the number of contacts each individual has (degree distribution), the extent to which those contacts are also connected to each other (clustering), and the average distance of those connections in a spatially distributed population (locality; spatial structure). This heterogeneity may substantially affect model predictions about the spatial spread of disease, infection/reinfection dynamics, local inter-strain competition and threshold behavior (May and Lloyd, 2001; Gupta and Hill, 1995; Pastor-Satorras and Vespignani, 2001; Meyers et al., 2003; Schinazi, 1999).

In areas where the burden of TB is low and continues to decline, localized outbreaks nonetheless sporadically occur. Variability in host susceptibility and strain-specific

*Corresponding author.
E-mail address: ccolijn@hsph.harvard.edu (C. Colijn).

0022-5193/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2007.04.015

differences in virulence and transmissibility (fitness) have been examined as explanatory factors for location-specific disease patterns (Valway et al., 1998; Murphy et al., 2002). Here we explore the null hypothesis that localized outbreaks can occur during declining epidemics as a result of locally constrained contact structure, even when the population is otherwise homogeneous. In order to test this hypothesis, we develop two models of TB epidemics that encapsulate the same natural history: a baseline differential equation model imposing homogeneous mixing, and a network model on a class of spatially structured networks. We modify the extent to which contacts are constrained to be local on the networks and examine declining epidemics under fully homogeneous mixing, networks with long-range contacts and networks with short-range contacts.

2. The models

2.1. Natural history of TB

The dynamics of TB within individual hosts (sometimes called the disease's *natural history*) are complex. Upon infection, individuals enter a latent state during which they are not infectious or symptomatic. From latency, there are three routes to active TB: primary progression, in which the infection progresses to active disease within the first 5 years; endogenous reactivation, in which an old infection activates, and exogenous reinfection, in which a new infection, acquired after an older infection, progresses to active disease.

The rate of progression from latency to active disease varies with the time since infection: for the first approximately 5 years after infection, this progression rate p_1 is considerably higher than it is subsequently (p_2) (Sutherland et al., 1976; Horwitz, 1969; Vynnycky and Fine, 1997). If an individual does not progress from latent infection to active disease during the first few years after the initial infection, he or she may remain latently infected for many years. However, a new exposure to the disease is thought to transiently increase the risk of progression.

The variable progression rate from latency to active disease has been modeled in several ways. Some modelers have split the latent class into "fast" and "slow" progressors, with fixed portions α and $1-\alpha$ of susceptibles entering fast-progressing and slow-progressing latent classes upon infection. This structure specifies that a portion $1-\alpha$ of individuals have some innate protection from TB, while a portion α are predestined to be "fast progressors". While this approach has the advantage of simplicity, it has the disadvantage that such models are very sensitive to α, which is very difficult to measure. Other authors have modeled this variable progression rate by including arbitrarily distributed latent periods (Feng et al., 2001) or using partial differential equations that include age and maturation of infection (Vynnycky and Fine, 1997).

Here, we present two models in which we avoid predestining the portion of fast and slow progressors. Both of the models are based on an SEIR framework where the recovered class, R, is accessible only after antibiotic treatment is introduced (circa 1950). Following Vynnycky and Fine (1997), we assume that the original infection confers some partial immunity which protects against progression to active disease, but not against the acquisition of a new infection.

While infants are at high risk of disease if infected with TB, because smear-positive pulmonary TB (the most infectious manifestation of disease) is rare in childhood, children are not thought to play an important role in the continued transmission of disease within communities (Styblo, 1991). Additionally, in areas where TB has been declining for many years, the average age of infection will be relatively high and there is likely only to be a small number of children who have been infected by TB. For these reasons, we have chosen to represent only adults in our model and have chosen baseline parameter values to represent the natural history of disease among these individuals. Because our intent is to examine the hypothesis that localized outbreaks can emerge in a homogeneous population with contact structure, we do not include sources of individual heterogeneity such as age-specific risks of progression, variable susceptibility or differences in TB strain transmissibility or virulence.

2.2. Modeling approach

Most models of TB epidemics have been differential equation models that assume a homogeneously mixed population. Our goal is to explore the effect of non-random mixing. To this end it is useful to have not only a network model, but also a baseline model that assumes homogeneous mixing and represents the natural history of TB in the same way. We therefore develop a differential equation model, making use of a delay to include the dependence of the risk of disease progression on the time since infection.

In the differential equation model, the population is homogeneously mixed, so it is not possible to examine local inhomogeneities. However, it is possible to directly measure the contribution of reinfection to disease incidence; if disease is locally clustered we expect to observe an increased frequency of reinfection. Therefore, comparing reinfection in the two models allows us to estimate the amount of additional reinfection induced by the introduction of contact structure. If local contact structure leads to substantially increased reinfection, this indicates that there is sufficient local clustering of disease to affect the dynamics of transmission, which has implications for policy control (Gomes et al., 2004).

In the network model, we can also directly estimate the amount of spatial variability in disease burden; this quantity does not have a directly comparable analogue in the differential model.

2.2.1. Delay differential equation model

In our differential equation model, susceptible individuals enter the fast-progressing latent stage L^1 upon infection. This stage lasts τ years ($\tau = 5$), and during this time individuals progress to infectious disease at rate p_1 and die at a rate μ. The number of individuals in the fast-progressing latent class at any time is given by

$$L^1(t) = \int_{t-\tau}^{t} \beta I S e^{-(\mu+p_1)(t-s)}\, ds, \quad (1)$$

where β is the transmission parameter, which is related to the contact rate and the per contact transmission rate, I is the fraction of the population that is infectious, S is the fraction susceptible and μ is the natural death rate. Upon leaving L_1, $\beta I_\tau S_\tau e^{-(\mu+p_1)\tau}$ enters the slower latency L_2. Meanwhile some portion has died at rate μ while in L_1, and some have progressed to active disease.

Those that progress are now in the infectious state, which gains a delayed term

$$P(t) = \beta I_\tau S_\tau e^{-\mu\tau}(1 - e^{-p_1\tau}).$$

We use subscripts to denote delays: $x_\tau = x(t-\tau)$. $P(t)$ is the contribution of primary progression to disease incidence. The same approach for the re-infected high-risk latency class generates a similar delayed contribution to the infectious class, where rather than entering from the susceptible class, individuals are re-infected from the latent and recovered classes. This gives the exogenous reinfection term

$$E(t) = \beta e^{-\mu\tau}(1 - e^{-p_r\tau}) I_\tau (L_\tau + R_\tau).$$

$E(t)$ is the contribution of reinfection to disease incidence. We have used the approximation that those who progress from fast-progressing latency to active disease all do so a time τ after infection. This results in the factor $1 - e^{-p_1\tau}$ and $1 - e^{-p_r\tau}$ in the expressions for $P(t)$ and $E(t)$. This limits the accuracy of the differential equation model, which is intended to describe the dynamics on time scales longer than $\tau = 5$ years. The main advantage to doing this is that we can readily compare the portion of new disease due to primary infection to that due to exogenous reinfection (as described in Section 2). This portion is not fixed by a predetermined fraction α of susceptibles that enter the fast-progressing latency, but arises from the time-varying progression rates.

If the disease progresses, the individual may die (at rate μ_{TB}), self-recover and return to the latent state (at rate r; this is a form of recovery that does not depend on treatment), or be treated and move to the recovered state (at rate r_{TR}). Recovered individuals may die (at rate μ), relapse at rate r_{rel} to the infectious state, or become re-infected.

This gives rise to the following set of delay differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= \gamma - \beta I S - \mu S,\\
\frac{dL}{dt} &= \beta e^{-(\mu+p_1)\tau} I_\tau S_\tau + \beta e^{-(\mu+p_r)\tau} I_\tau (L_\tau + R_\tau)\\
&\quad - \beta I L - p_2 L - \mu L + rI,\\
\frac{dI}{dt} &= \beta e^{-\mu\tau}(1 - e^{-p_1\tau}) I_\tau S_\tau + \beta e^{-\mu\tau}(1 - e^{-p_r\tau}) I_\tau (L_\tau + R_\tau)\\
&\quad + p_2 L + r_{rel} R - rI - \mu_{TB} I - r_{TR} I\\
\frac{dR}{dt} &= r_{TR} I_{\tau_2} - r_{rel} R - \beta I R - \mu R, \quad (2)
\end{aligned}$$

where S is the fraction of the population in the susceptible state, L is the fraction in the slow-progressing latent state, I is the fraction that are infectious, R is the fraction recovered, and the parameters are as defined in Table 1. We ensure a constant population by setting

$$\gamma = \mu(1 - I) + \mu_{TB} I. \quad (3)$$

2.2.2. Network model

In the network model, individuals are represented by vertices of the network and transmission of disease may occur between two individuals only if there is an edge connecting the corresponding vertices. We place the vertices randomly with uniform distribution on a square patch and choose whether or not to connect two vertices based on their locations. This gives a two-dimensional spatial structure to the network. We tune how much preference there is for edges that connect individuals that are near each other so that we can choose to have either locally constrained contacts or more long-range contacts.

Since the spatial density of the network is uniform, there are always fewer nearby vertices than far away ones. Thus, when we choose to make contacts local, groups of vertices that are near each other tend to be connected in clusters. These networks have higher clustering coefficients than networks with long-range contacts: on local networks, the contacts of a given individual are likely also to be contacts of each other.

The preference for short or long-range contacts is specified using a spatial kernel, in which an edge between two vertices i and j is formed with probability

$$p = \frac{n}{2\pi D^2} e^{-d_{ij}^2/2D^2}. \quad (4)$$

Here, d_{ij} is the distance between the two vertices, D tunes the "locality" of the network, and n is the average degree of the resulting graph. (In practice due to finite populations, the average degree may be somewhat less than n.) This class of graphs has been used in disease models previously (Read and Keeling, 2003). As D increases, the preference for short edges decreases, and the graph contains more long-range contacts. However, when D is low, most of the contacts of a given individual are near that individual. The degree distribution is Poisson with mean n, regardless of D. The sociological connection to this approach is that low D networks reflect populations with clustered social groups, in which members are in contact with each other and are likely to know

Table 1
Model parameters

Parameter	Description	Value	Unit	Source
Both models				
μ	Natural death rate	0.02	year^{-1}	Cohen and Murray (2004)
μ_{TB}	TB mortality rate	0.3	year^{-1}	Dye et al. (1998), Springett (1971)
f	Partial immunity	0.4	none	Dye et al. (1998), Sutherland (1976)
r_{rel}	Relapse rate	0.05	year^{-1}	Blower et al. (1995), Cohen et al. (2007)
r	Self-recovery rate	0.2	year^{-1}	Dye et al. (1998)
r_{TR}	Treatment rate	0.9	year^{-1}	Assumed
f_T	Portion with treatment	0.9	None	Assumed
p_1	Primary progression rate	0.03	year^{-1}	Vynnycky and Fine (1999), Dye et al. (1998)
p_2	Endogenous activation rate	0.0003	year^{-1}	Vynnycky and Fine (1999)
f_{imm}	Partial immunity	0.4	None	Vynnycky and Fine (1999)
p_r	$(1-f_{imm})p_1$	0.018	year^{-1}	Vynnycky and Fine (1999)
Delay model				
τ	Duration of fast latency	5	year	Vynnycky and Fine (1999)
β	Transmission parameter before 1900	10	year^{-1}	Calculated
β_2	Transmission parameter after 1900	6	year^{-1}	Calculated
Network model				
B	Transmission parameter before 1900	0.9	$\text{year}^{-1}\,\text{contact}^{-1}$	Fitted
B_2	Transmission parameter after 1900	0.5	$\text{year}^{-1}\,\text{contact}^{-1}$	Fitted
c	Average number of contacts	15	No unit	Assumed

each other's friends, colleagues and families. High D networks would represent populations where people are less likely to belong to such social groups.

The natural history of TB in the network model is the same as in the differential equation model: a susceptible individual in contact with k infectious individuals has a probability of becoming infected of $p = 1-(1-B)^k$ in a unit of time, where B is the per contact transmission probability per unit time. Upon infection, the individual is moved to the latent class and there suffers a progression rate p_1 to active disease. After 5 years, this rate is decreased to $p_2 \ll p_1$. The individual is then subject to a reinfection probability per unit time $1-(1-B)^k$ (given contact with k infectious individuals), and for the next 5 years, the progression probability is p_r, where $1-p_r/p_1$ is the partial immunity conferred by the first infection. After those 5 years, the progression rate is reset to p_2. Other transitions are as described above.

The birth rate is set to keep the population stable, but not fixed. Births may replace dead individuals. If this occurs for a vertex x in the "dead" state, the new individual is not simply placed at the location of x. Rather, a neighbor y of x moves into its place, and a neighbor z of y moves into y's location, and so on. After a specified number of replacements, a susceptible is inserted into the network. In this way, we avoid introducing spatial correlations between death due to disease and birth of new susceptibles without incurring the computational costs associated with regeneration of the network each time a new vertex is added.

2.3. Data

We are interested in the emergence of heterogeneity in declining epidemics in homogeneous populations to explore the null hypothesis that contact structure induces heterogeneity. For this reason we have not included HIV and drug resistance in the models as both would introduce individual-level heterogeneity. The data that are relevant for this study are from declining epidemics during periods in which HIV and drug-resistance have not had substantial effects on TB dynamics. Serial surveys of the annual risk of infection conducted in the Netherlands between 1910 and 1966 demonstrate a steady decline in transmission which accelerated after the introduction of TB antibiotics (Styblo et al., 1969). Additionally, Styblo reported a consistent 1:2:4 ratio of TB mortality:incidence:prevalence prior to the advent of antibiotic treatment (Styblo, 1991); we use these data and ratios, in combination with TB notification data from repeated surveys at 5 and 10 year intervals from Germany (DGDDR, 1980) and the Netherlands (Styblo, 1991) from the period 1951–1979 to characterize the general relationships between risk of infection, incidence and prevalence during declining TB epidemics.

We begin by allowing the model to equilibrate at approximate pre-1900 conditions; the comparison between this steady state and data from Styblo (1991) provides confirmation that our natural history parameters are reasonable and gives a baseline starting point for simulating declining disease levels. From this point, TB epidemics were observed to be in decline even before antibiotic treatment became available (Styblo, 1991). To mimic this

decline, we reduce the transmission parameter so that by 1950 a new equilibrium of TB incidence and prevalence is reached. In 1950 we then simulate the introduction of antibiotics, which functioned to decrease the duration of disease among those receiving treatment. For the spatial model, the epidemic dynamics differ depending on how locally constrained the contacts are, i.e. depending on the value D assumed for the contact network. Here, we choose a set of baseline parameters for an intermediate value of D (5) such that the resultant epidemic trajectories from both spatial and delayed differential equation models follow similar paths as observed declining epidemics. While the natural history parameters are derived from the clinical and epidemiological literature (see Appendix), the transmission parameter was adjusted so that the annual risk of TB infection, TB prevalence and TB incidence declined in a manner that reflects observed patterns in the Netherlands and Germany. Fig. 1 shows modeled incidence during declining epidemics plotted with data from the Netherlands and Germany during the era of antibiotic treatment.

3. Results

3.1. Dynamics of the delayed model

The long-term dynamics of the differential equation model are governed by an epidemic threshold R_0, such that when $R_0<1$ the disease-free equilibrium $(S,E,I,R)=(1,0,0,0)$ is stable and when $R_0>1$ it is unstable, and the endemic equilibrium is stable. In our case, R_0 is given by

$$R_0=\beta\left(\frac{p_1}{a_1b_1}(1-e^{-b_1\tau})-\frac{p_1}{a_1b_2}(1-e^{-(a_1+b_1)\tau})-\frac{p_2}{a_2b_2}e^{-(a_2+b_2+p_1-p_2)\tau}+\frac{p_2}{a_2b_1}e^{-(b_1+p_1-p_2)\tau}\right), \quad (5)$$

where $a_1=\mu_{TB}+r-p_1$, $a_2=\mu+r-p_2$, $b_1=\mu+p_1$ and $b_2=\mu_{TB}+r$. R_0 is proportional to the transmission parameter β and is independent of reinfection. R_0 given in Eq. (5) agrees with the result for R_0 in Feng et al. (2001) for a model without reinfection. The reason that reinfection does not play a role in the location of the $R_0=1$ threshold is that near the bifurcation, by definition the system is near the point $S=1$, $L=I=R=0$, where $S\gg L+R$ so that the contribution of reinfection to disease is negligible. Consistent with the findings of several previous modelers (Feng et al., 2000; Singer and Kirschner, 2004), an endemic equilibrium exists in the model even for $R_0<1$ when the reinfection progression rate p_r is sufficiently large (i.e. there is a backward bifurcation reflecting the fact that reinfection parameters do not occur in the expression for R_0). However, for this equilibrium to exist, p_r must be so large as to be unrealistic, i.e. even higher than p_1, the primary progression rate.

Even when $R_0<1$ and the disease-free equilibrium is stable, the approach to that equilibrium is slow. Fig. 2 shows the numerically determined real parts of the largest eigenvalues λ for each equilibrium, as a function of the transmission parameter β. This analysis was done with the DDE-BIFTOOLS package (Engelborghs et al., 2002, 2001). Near the equilibrium point, the transient approach to equilibrium has a time scale of $1/\lambda$. For the disease-free equilibrium the slope of λ near $\lambda=0$ is very small; even decreasing β, and hence R_0, to half its critical value, the time scale of approach is still about 80 years. This is consistent with the doubling times estimated in Blower et al. (1995). The resulting long transients, if they occur in real systems, present a challenge for evaluating policy interventions, because reliable indications of the interventions' effects would not be observable for many years. Long transient times indicate that the near-threshold

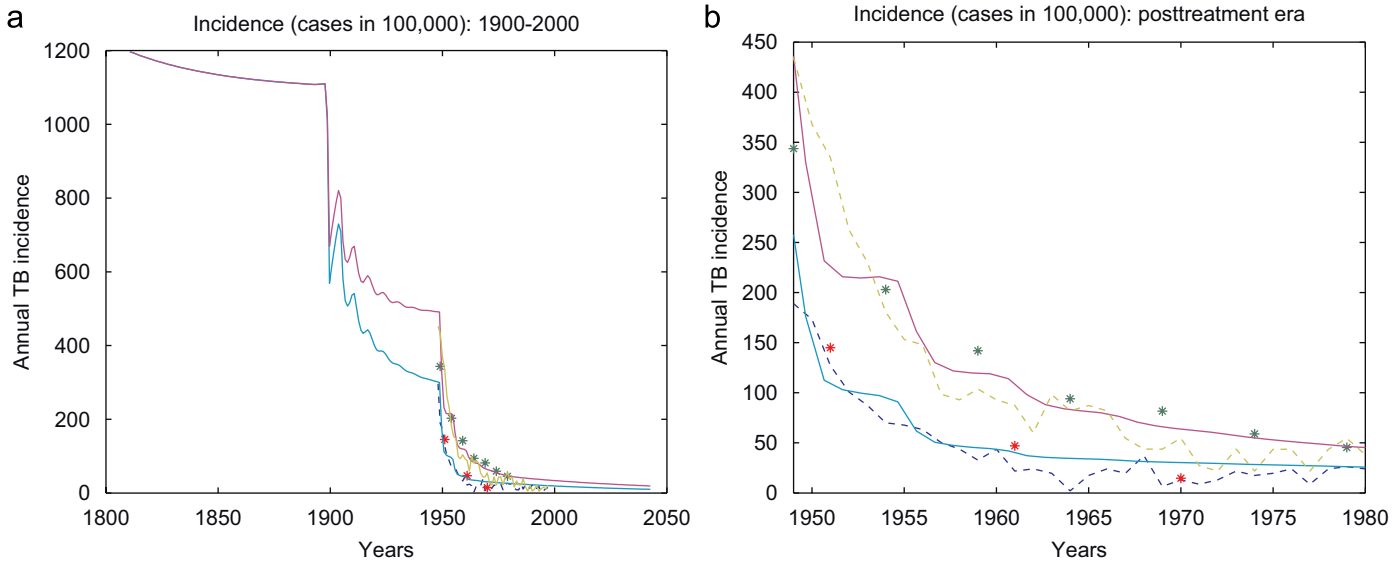

Fig. 1. The spatial and delay DE models with observed data for incidence after 1950. Dashed lines are the spatial model and solid lines are the delay model. (a) Incidence (cases in 100,000): 1900–2000. (b) Incidence (cases in 100,000): post-treatment era.

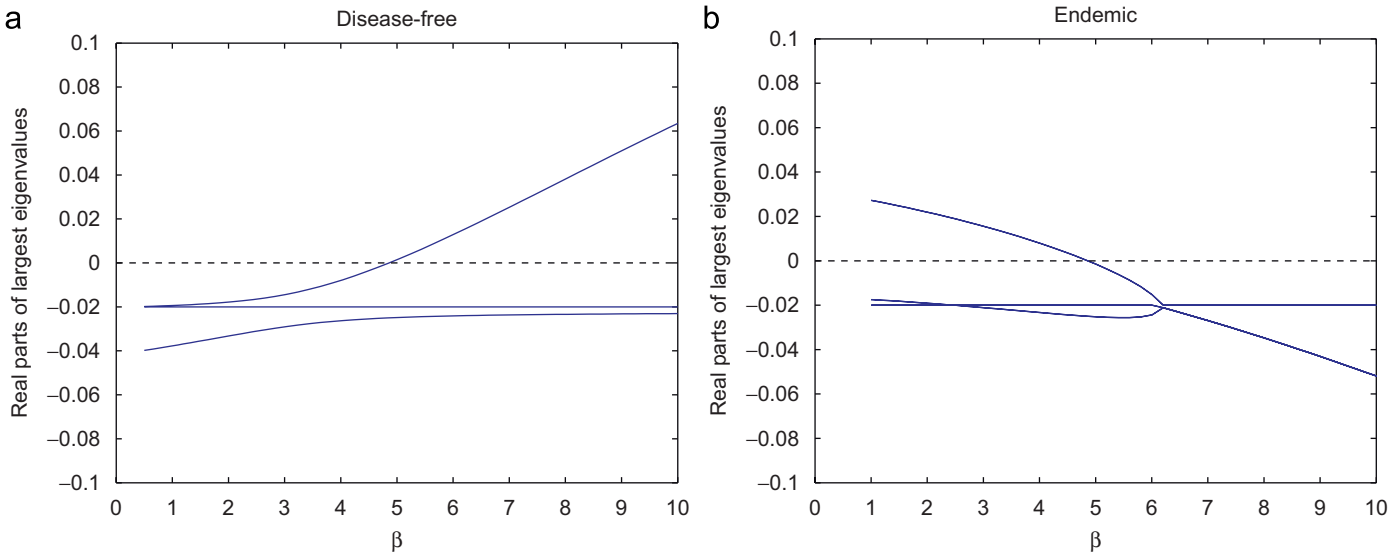

Fig. 2. Eigenvalues showing stability switch. (a) Disease-free. (b) Endemic.

declining phase of the epidemic, which is the phenomenon of interest here, will be of long duration.

The model's construction, in particular the form of the equation for $\dot{I}$, allows us to examine the relationship between primary progression to disease and exogenous reinfection. Consensus based on homogeneously mixed models such as this one has been that at low levels of disease reinfection occurs so rarely that it is negligible. Since we can expect reinfection to be higher in local outbreaks, comparing reinfection in the differential equation model and in the network model yields an estimate of how important localization of disease is for transmission dynamics.

From Eq. (2), we can examine the relative importance of primary progression and reinfection at steady state. To do this, we solve the steady-state equations for the ratio $\rho \equiv E/P$ of the contribution to $\dot{I}$ from exogenous reinfection to that from primary activation:

$$\rho = \frac{(1-\mathrm{e}^{-p_r\tau})(L_\tau + R_\tau)}{(1-\mathrm{e}^{-p_1\tau})S_\tau} \equiv \frac{c_2}{c_1}\frac{(L_\tau + R_\tau)}{S_\tau}. \tag{6}$$

At steady state the delayed arguments are equal to their steady-state values, so we may drop the subscripts, and now use I to refer to the steady-state value of $I(t)$, and likewise for S, L and R.

To get an expression for the ratio ρ, we add the second and third equations of Eq. (2) together, and use the first and fourth equations of (2) to find steady-state values of S and R in terms of the steady-state value of I. Let $(r_l+\mu)/\beta = a$, and we have

$$\begin{aligned}\frac{L+R}{S} = {} & \frac{I(\beta I+\mu)}{\gamma(1-\mathrm{e}^{-\mu\tau})\beta I+\mu} \\ & \times\left(\mathrm{e}^{-\mu\tau}r_{TR}\frac{I}{I+a}+\frac{r_l r_{TR}}{\beta}\frac{1}{I+a}-(\mu_{TB}+r_{TR})\right) \\ & +\frac{\mathrm{e}^{-\mu\tau}\beta I}{(1-\mathrm{e}^{-\mu\tau})\beta I+\mu}+\frac{r_{TR}}{\beta\gamma}\frac{I}{I+a}(\beta I+\mu).\end{aligned} \tag{7}$$

Since $\mu\tau \sim 0.1$ we can approximate $\mathrm{e}^{-\mu\tau} \approx 1-\mu\tau$. Then Eq. (7) becomes

$$\begin{aligned}\frac{L+R}{S} = {} & \frac{I}{\gamma}\frac{I}{I+1/\tau\beta}\left(\frac{I}{\mu\tau}+\frac{1}{\tau\beta}\right) \\ & \times\left(\mathrm{e}^{-\mu\tau}r_{TR}\frac{I}{I+a}+\frac{r_l r_{TR}}{\beta}\frac{1}{I+a}-b\right) \\ & +\frac{((1-\mu\tau)/\mu\tau)I}{I+1/\tau\beta}+\frac{r_{TR}}{\beta\gamma}\frac{I}{I+a}(\beta I+\mu).\end{aligned} \tag{8}$$

The realistic steady-state values of I (namely the observed prevalence of active TB in the world) are very low: a prevalence of 0.02, or 2000 cases in 100,000, is considered a high TB burden. This motivates rescaling I. Also, since β, the transmission parameter, is the least well-known parameter in the system, it is useful to group other parameters together but retain the dependence on β. With these two points in mind, we write $x = I/\mu$, $\alpha_1 = 1/(\tau\mu)$, $\alpha_2 = (r_l+\mu)/\mu$, and rewrite Eq. (7) in terms of x:

$$\begin{aligned}\rho = {} & \frac{c_2}{c_1}\left\{\frac{r_{TR}}{\tau\gamma}\left(\mathrm{e}^{-\mu\tau}\beta+\frac{r_l}{\mu}\right)\left(\frac{\beta x+1}{\beta x+\alpha_1}\right)\left(\frac{x}{\beta x+\alpha_2}\right)\right. \\ & -\left(\frac{\mu_{TB}+r_{TR}}{\tau\gamma}\right)x\left(\frac{\beta x+1}{\beta x+\alpha_1}\right) \\ & \left.+\frac{1-\mu\tau}{\mu\tau}\left(\frac{\beta x}{\beta x+\alpha_1}\right)+\frac{\mu r_{TR}}{\gamma}x\left(\frac{\beta x+1}{\beta x+\alpha_2}\right)\right\}.\end{aligned} \tag{9}$$

Naturally, changing parameters changes the steady-state value of I and therefore x, so one might question whether Eq. (9) provides any understanding of the relationship between β and the amount of reinfection. In fact, the only place that the progression rates p_1 and p_r occur in Eq. (9) are in the ratio c_2/c_1, and furthermore, I and hence x are very sensitive to these rates. Therefore, the steady-state value of x *can* be changed without changing β and without altering the relative contributions of the different terms in Eq. (9).

The function ρ is $O(x)$ as $x \to 0$. This is consistent with the biological literature on reinfection in TB; it is generally thought that as disease prevalence decreases, reinfection is not a significant contributor to disease incidence. This view is largely informed by the assumption of homogeneous population mixing either in qualitative discussions or in previous models. A plot of ρ is shown in Fig. 3 for several value of β. Note that in the DDE model, exogenous reinfection is as important as primary infection for disease prevalence greater than approximately 450 cases in 100,000. This is considered to be a high TB burden (WHO, 2006). Previous work has suggested that there may be a threshold beyond which reinfection accounts for much of the prevalence of disease (Gomes et al., 2004); in our delay model we do not see a threshold but reinfection does increase monotonically with the transmission parameter. In contrast, the network model with localized contacts predicts a much higher contribution of reinfection (also shown in Fig. 3).

In the differential equation model, disease elimination is possible, but the approaches to both the disease-free and the endemic equilibrium are very long, so that behavior observed during declining epidemics can be expected to persist for extended periods of time. The model predicts that at low disease levels there will be some reinfection, but the ratio of incidence from reinfection to incidence from primary infection decreases gradually with disease prevalence, approaching zero as disease is eliminated.

3.2. Dynamics of the network model

In the network model, when D is low (local networks), disease levels are lower than those predicted both by the differential equation model and by higher D networks. However, reinfection contributes notably more to disease. Indeed, the effect of network structure on reinfection is quite marked; Fig. 3 shows that as disease prevalence decreases to zero, the contribution of reinfection to incidence remains substantial.

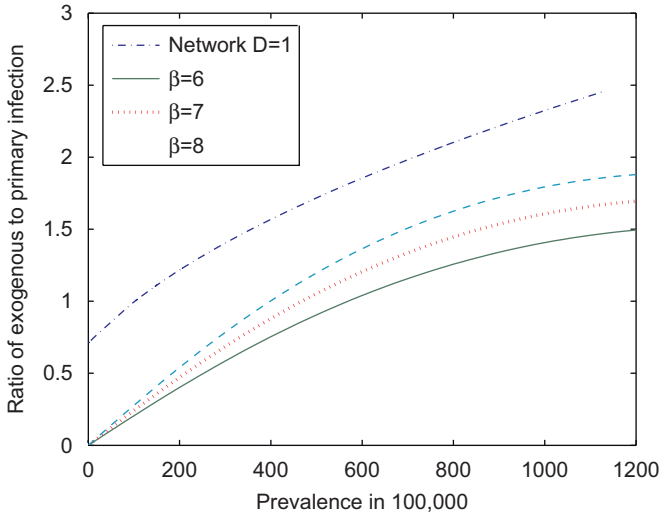

Fig. 3. The ratio of exogenous reinfection to primary activation for the network model with $D = 1$ and the differential equation model for several values of the transmission parameter, β.

The higher levels of reinfection in low D networks occur because when contacts are more local, the network has a higher clustering coefficient. Infection will follow a random walk beginning at an infected vertex; as the clustering coefficient increases, so does the probability that such a random walk will intersect itself, causing reinfection. While clustering increases the probability of reinfection, it decreases the overall spread of disease, because distant parts of the network are less likely to be reached so that some pockets of susceptible individuals are isolated and thereby protected.

The network model with long-range contacts (high D) has average disease trajectories similar to those of the differential equation model; trajectories approach the differential equation trajectories as D rises, due to the increased rate of spatial diffusion in such networks.

In any spatial model of disease transmission, the force of infection experienced by individuals can vary such that the local risk of infection can be high even when the average risk of infection is low. In the most obvious case, this occurs if we introduce one or a few infectious individuals into an otherwise susceptible population where most contacts are local. Spatial clustering in the resulting rising epidemics would be expected whenever there are local constraints on the contact structure.

In contrast, we observe a heterogeneous force of infection during *declining* epidemics when the initial conditions are homogeneous. We populate the spatial model with a uniformly distributed number of latently infected and infectious individuals (mimicking pre-1900 conditions) and allow the disease to decline on a relatively localized network (low D). Both latency and disease become spatially clustered as the epidemic recedes. At any point in time these clusters appear as localized outbreaks of disease. They emerge from homogeneous initial conditions as a result of the network structure and the decline of the epidemic.

The differential equation model has no direct analogue for the variation in local disease burden, as it has no local structure. However, we can quantify the difference between the variation in local disease burden that we observe, and what we would expect if the spatial distribution of latency remained uniform. To do this we examine the variance in local prevalence of latent infection during the declining epidemic.

Let the set of incident vertices of a vertex x be denoted $A(x)$, and the subset that are in state M be $A_M(x)$. Let p_m be the portion of the whole population in state M. We can determine what the variance of $|A_M(x)|$ would be if there were no spatial correlations or additional structure on the graph—i.e. if the different states were uniformly distributed. The probability that a given vertex x has n neighbors

is given by the Poisson distribution

$$p(n) = \frac{v^n e^{-v}}{n!},$$

where v is the average degree. If there are no spatial correlations or additional structure, then every vertex has an equal probability p_m of being in state M, and we have

$$P(|A_M(x)| = k) = \sum_{n=k}^{\infty} \frac{v^n e^{-v}}{n!} \binom{n}{k} p_m^k (1-p_m)^{n-k}.$$

The second moment, and hence the variance, of $|A_M(x)|$, can be found by writing the second moment as

$$m_2 = \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} \frac{v^n e^{-v}}{n!} k^2 p_m^k (1-p_m)^{n-k}$$
$$= e^{-v} \sum_{i=0}^{\infty} \frac{i^2 p_m^i v^1}{i!} \sum_{j=0}^{\infty} \frac{(1-p_m)^j v^j}{j!}$$

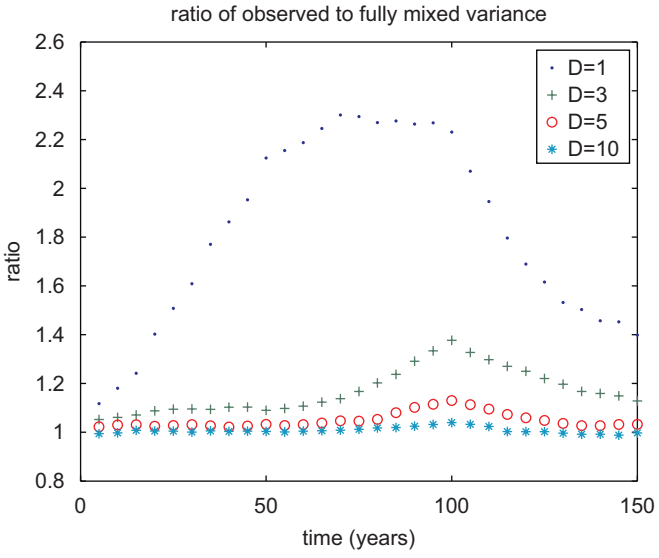

Fig. 4. The variance of local latency compared to the predicted variance without spatial clustering.

which can be written as

$$m_2 = e^{-p_m v} \left(v \frac{d}{dv} \right)^2 e^{p_m v} = p_m v (1 + p_m v). \quad (10)$$

The mean is $p_m v$, so the variance is $m_2 - p_m^2 v^2 = p_m v$. This is the variance that we should see if every vertex has the same probability p_m of being in state M.

Fig. 4 shows the ratio of the actual variance of $|A_M(x)|$, observed during epidemics on the network, to $p_m v$, where M is the latent state, along the epidemic trajectories with graphs with different D values. The variances are higher than would be predicted by Eq. (10), and near the middle of the $D = 1$ trajectory the actual variance is more than twice what it would be without spatial structure. In other words, in declining epidemics on local networks, a substantial number of individuals see much higher local levels of infection than they would if mixing were random.

Fig. 5 shows the evolving distribution of the local prevalence of latency along a realistically declining TB epidemic, on networks with $D = 1$ and 10. The plots illustrate the qualitative differences between the local and global networks. The plot for $D = 1$ shows that skew-symmetry develops so that a larger number of the vertices see a locally high disease burden than on the $D = 10$ networks where no skew-symmetry is evident.

4. Conclusion

Our results indicate that clustering of disease can emerge during declining TB epidemics without the explicit inclusion of host or strain differences in susceptibility and fitness. Thus, some local outbreaks can be expected to occur even in the absense of these sources of individual heterogeneity, simply as a result of the contact structure of the population. This phenomenon is not specific to the data we chose for parameterizing the model, but occurs in a wide range of declining epidemics with different rates of decline, starting points and parameter values. Interestingly,

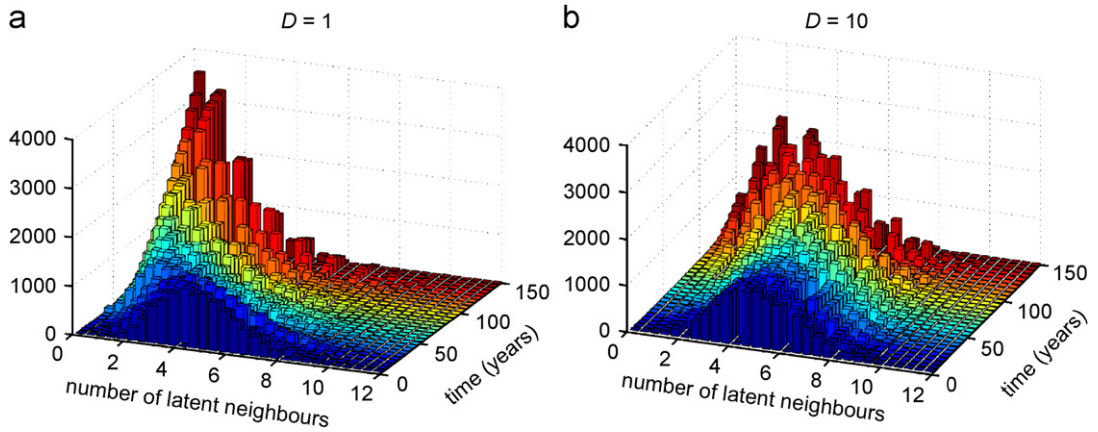

Fig. 5. The time evolution of local latency for $D = 1$ and 10. Height indicates the number of individuals with the given number of latent neighbors at the corresponding time. At 50 years we decrease the transmission parameter to mimic the decline in TB between 1900 and 1950, and at 100 years (1950), we introduce antibiotic treatment. (a) $D = 1$. (b) $D = 10$.

inhomogeneity arises spontaneously from uniformly distributed initial states. Of course, in real populations, both host and strain differences do exist. Our findings do not discount their likely importance, but rather suggest that non-random mixing also has a substantial effect on clustering of disease in low incidence areas.

The increase in the variance of local disease prevalence indicates that the epidemic may be near the model's phase transition, i.e. near the boundary between the disease-free absorbing state and the endemic equilibrium. This hypothesis is supported by differential equation models (Salpeter and Salpeter, 1998), including ours. In a phase transition, spatial structure across a range of scales is typical, and we would expect local outbreaks with a wide range of sizes, even in the absence of strain or host heterogeneity. While a thorough examination of the phase transition is beyond the scope of the current paper, the characterization of phase transitions and critical behavior in TB models is a promising avenue for further research.

At parameter values that are reasonable for TB epidemics, the time scale of the approach to the disease-free equilibrium in the differential equation model is very long. We also observe long time scales in the network model, particularly for local networks, in which there is slower spatial diffusion. This means that the phase transition that occurs as the epidemic crosses the threshold would last for an extended period of time. This in turn indicates we would expect to observe the spontaneous emergence of spatial structure over long periods during declining epidemics.

Contact structure has a substantial impact on reinfection. In the differential equation model the importance of exogenous reinfection falls to zero as disease prevalence decreases. But in networks with local contact structure, the ratio of reinfection to primary infection does not approach zero as disease prevalence falls. Even when the overall prevalence of disease is very low, sufficient spatial structure (i.e. pockets of relatively high latency) emerges on these networks that reinfection is not negligible. In local outbreaks with significant reinfection, policies aimed at preventing progression from primary infection (such as one-time prophylactic treatment with isoniazid) will be less effective because individuals who have been exposed once are more likely than others to be exposed again. In addition, in areas of high local prevalence of latency, if an individual in a closely linked clustered community acquires drug-resistant disease through inadequate treatment, that individual may re-infect his or her contacts, who remain susceptible to this new strain even if they received prophylactic treatment after a previous exposure to TB. Since their subsequent exposures would not be detected, they would not receive prophylaxis again, and would therefore be susceptible to disease with the new strain. In this manner, reinfection in local high-burden pockets can facilitate the spread of drug-resistant strains.

In summary, non-random mixing in a population leads to the emergence of local disease clusters in declining TB epidemics. This occurs even when the initial conditions are spatially homogeneous and individual heterogeneity (beyond that imposed by non-random mixing) is excluded. In these local clusters, disease is sufficiently concentrated that reinfection becomes a significant contributor to overall TB levels. This effect on the disease transmission dynamics has implications for control policy and the emergence of new strains.

Appendix A. Parameter estimation

The parameters used in our models are adopted from the clinical literature and are consistent with values used in previous models of tuberculosis.[1] In some cases, there has been substantial variability in the value for specific parameters used previously (e.g. the rate of progression from infection to disease (Blower et al., 1995; Dye et al., 1998). In these situations we have chosen a value that corresponds better with observational data. The value of the parameter corresponding to the duration of fast latency (τ) depends on the time during which one assumes that an individual is at elevated risk following a recent infection; here, we have adopted the convention used by Vynnycky and Fine (1997). As emphasized in the manuscript, the pattern of connections in the network model affects disease spread and contraction. For the purpose of this preliminary analysis we have set the average number of close respiratory contacts to be 15. In reality, the structure of connections is likely to be more dynamic than that represented here and new empiric studies on mixing patterns and the transmission of infectious diseases will inform future work (Edmunds et al., 2006).

References

Blower, S.M., McLean, A.R., Porco, T.C., Small, P.M., Hopewell, P.C., Sanchez, M.A., Moss, A.R., 1995. The intrinsic transmission dynamics of tuberculosis epidemics. Nat. Med. 1 (8), 815–821.

Blower, S.M., Small, P.M., Hopewell, P.C., 1996. Control strategies for tuberculosis epidemics: new models for old problems. Science 273 (5274), 497–500.

Cohen, T., Murray, M., 2004. Modeling epidemics of multidrug-resistant *M. tuberculosis* of heterogeneous fitness. Nat. Med. 10 (10), 1117–1121.

Cohen, T., Lipsitch, M., Walensky, R.P., Murray, M., 2006. Beneficial and perverse effects of isoniazid preventive therapy for latent tuberculosis infection in HIV-tuberculosis coinfected populations. Proc. Natl. Acad. Sci. USA 103 (18), 7042–7047.

Cohen, T., Colijn, C., Finklea, B., Murray, M., 2007. Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission. R. Soc. Interfaces 4, 523–531.

DGDDR, 1980. Das Gesundheitswesen der Deutschen Demokratischen Republik. Jargang, Berlin.

Dye, C., 2006. Global epidemiology of tuberculosis. Lancet 367 (9514), 938–940.

[1]Note that since the differential equation is the $\Delta t \to 0$ limit of the discrete time stochastic system, rates in the DE model are given in terms of their corresponding discrete values by $r_{de} = -\ln(1 - r_{st})$. Values listed in Table 1 are the stochastic values.

Dye, C., Garnett, G.P., Sleeman, K., Williams, B.G., 1998. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. Directly observed short-course therapy. Lancet 352 (9144), 1886–1891.

Edmunds, W.J., Kafatos, G., Wallinga, J., Mossong, J., 2006. Mixing patterns and the spread of close-contact infectious diseases. Emerg Themes Epidemiol 3 (1), 10.

Engelborghs, K., Luzyanina, T., Samaey, G., 2001. DDE-BIFTOOL v. 2.00: a Matlab package for bifurcation analysis of delay differential equations. Technical report, Department of Computer Science, K.U.Leuven, Leuven, Belgium.

Engelborghs, K., Luzyanina, T., Roose, D., 2002. Numerical bifurcation analysis of delay differential equations using DDE-BIFTOOL. ACM Trans. Math. Software, 1–21.

Feng, Z., Castillo-Chavez, C., Capurro, A.F., 2000. A model for tuberculosis with exogenous reinfection. Theor. Popul. Biol. 57 (3), 235–247.

Feng, Z., Huang, W., Castillo-Chavez, C., 2001. On the role of variable latent periods in mathematical models for tuberculosis. J. Dyn. Differential Equations 13 (2), 425–452.

Gomes, M.G.M., Franco, A.O., Gomes, M.C., Medley, G.F., 2004. The reinfection threshold promotes variability in tuberculosis epidemiology and vaccine efficacy. Proc. Biol. Sci. 271 (1539), 617–623.

Gupta, S., Hill, A.V., 1995. Dynamic interactions in malaria: host heterogeneity meets parasite polymorphism. Proc. Biol. Sci. 261 (1362), 271–277.

Horwitz, O., 1969. Public health aspects of relapsing tuberculosis. Am. Rev. Respir. Dis. 99 (2), 183–193.

May, R.M., Lloyd, A.L., 2001. Infection dynamics on scale-free networks. Phys. Rev. E 64 (6), 066112.

Meyers, L., Newman, M., Martin, M., Schrag, S., 2003. Applying network theory to epidemics: control measures for *Mycoplasma pneumoniae* outbreaks. Emerging Infect. Dis. 9 (2), 205.

Murphy, B., Singer, B., Anderson, S., Kirschner, D., 2002. Comparing epidemic tuberculosis in demographically distinct heterogeneous populations. Math. Biosci. 180, 161–185.

Pastor-Satorras, R., Vespignani, A., 2001. Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86 (14), 3200–3203.

Read, J.M., Keeling, M.J., 2003. Disease evolution on networks: the role of contact structure. Proc. Biol. Sci. 270 (15), 699–708.

Salpeter, E.E., Salpeter, S.R., 1998. Mathematical model for the epidemiology of tuberculosis with estimates of the reproductive number and infection-delay function. Am. J. Epidemiol. 147 (4), 398–406.

Schinazi, R.B., 1999. On the spread of drug-resistant diseases. J. Stat. Phys. 97 (1–2), 409–417.

Singer, B., Kirschner, D., 2004. Infuence of backward bifurcation on interpretation of R_0 in a model of epidemic tuberculosis with reinfection. Math. Biosci. Eng. 1 (1).

Springett, V.H., 1971. Ten-year results during the introduction of chemotherapy for tuberculosis. Tubercle 52 (2), 73–87.

Styblo, K., 1991. Epidemiology of tuberculosis. Royal Netherlands Tuberculosis Association (KNCV).

Styblo, K., Meijer, J., Sutherland, I., 1969. Tuberculosis Surveillance Research Unit Report No. 1: the transmission of tubercle bacilli; its trend in a human population. Bull. Int. Union Tuberc. 42, 1–104.

Sutherland, I., 1976. Recent studies in the epidemiology of tuberculosis based on the risk of being infected with tubercle bacilli. Adv. Tuberc. Res. 19, 1–63.

Sutherland, I., Svandova, E., Radhakrishna, S., 1976. Alternative models for the development of tuberculosis disease following infection with tubercle bacilli. Bull. Int. Union Tuberc. 51 (1), 171–179.

Sutherland, I., Svandová, E., Radhakrishna, S., 1982. The development of clinical tuberculosis following infection with tubercle bacilli. 1. A theoretical model for the development of clinical tuberculosis following infection, linking from data on the risk of tuberculous infection and the incidence of clinical tuberculosis in the Netherlands. Tuberc. 63 (4), 255–268.

Valway, S., Sanchez, M., Shinnick, T., Orme, I., Agerton, T., Hoy, D., Jones, J., Westmoreland, H., Onorato, I., 1998. An outbreak involving extensive transmission of a virulent strain of *Mycobacterium tuberculosis*. N. Engl. J. Med. 338 (10), 633–639.

Vynnycky, E., Fine, P.E., 1997. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol. Infect. 119 (2), 183–201.

Vynnycky, E., Fine, P.E., 1999. Interpreting the decline in tuberculosis: the role of secular trends in effective contact. Int. J. Epidemiol. 28 (2), 327–334.

WHO, K., 2006. Global tuberculosis control—surveillance, planning, financing. World Health Organization, Geneva.