On the Effect of Population Heterogeneity on Dynamics of Epidemic Diseases[1]

Vladimir M. Veliov[2]

Abstract

The effect of the heterogeneity on the evolution of an epidemic disease in a heterogeneous population is investigated within a simple distributed parameter model. It is established that there exist short run and long run threshold values for the initial "weighted prevalence", at which the dependence of the short run (long run, respectively) behaviour of the disease on the initial level of heterogeneity changes qualitatively. The threshold value for the short run behaviour turns out to be independent of the particular data and equals 0.5.

The paper contributes also to the issue of modeling the dynamics of the disease in a heterogeneous population by non-distributed equations. A suitable class of simple functions is proposed that can be used as approximations of the "prevalence–to–incidence" function in models where the heterogeneity is not explicitly involved.

Keywords: epidemic disease, population heterogeneity, incidence function

1 Introduction

Epidemic models that explicitly take into account the heterogeneity of a population involve distributed parameter systems, see e.g. Diekmann, Heesterbeek, and Metz [4], Coutinho at al. [2], Diekmann and Heesterbeek [3]. Such models are not only more complex for numerical processing, but require distributed data that are usually not available. For this reason it is desirable to deal with non-distributed (aggregated) models, usually obtained by an appropriate averaging. In the present paper, starting from a rather general distributed SIS system modeling the dynamic of an epidemic disease in a heterogeneous population, we pass to a non-distributed system whose solution coincides with the aggregated solution of the distributed system. This ordinary differential system is not explicitly defined, but we establish that it represents an ordinary SIS model in a homogeneous population which,

[1] This research was partly supported by the Austrian Science Foundation under contract N0. 15618-OEK and by the Austrian National Bank (ÖNB) under grant N0. 9414

[2] Institute for Econometrics, Operations Research and Systems Theory, Vienna University of Technology, Argentinierstrasse 8/119, A-1040 Vienna, Austria, and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria, e-mail: vveliov@eos.tuwien.ac.at

however, involves a nonlinear "prevalence-to-incidence" function as a multiplicative component of the infection rate. We establish some properties of this nonlinear incidence function and propose a class of approximating functions depending on three parameters only, which could be identified by a modest amount of measurements of the prevalence. The analysis is motivated mainly by HIV/AIDS and Hepatitis, but could be useful also for other diseases.

We compare two populations with the same parameters (relevant to the considered model), and the same mean value of risk, the two populations differing only in the distribution of the individuals along the risk scale: one of the populations is more heterogeneous with respect to the risky behaviour than the other. We establish that there is a threshold value such that less heterogeneity is more advantageous in a short run if the initial *weighted prevalence* is below this threshold, and less advantageous in the opposite case. In addition, this threshold value turns out to be independent of data and equals 0.5. We establish a similar property also for the long run (asymptotic) behaviour, where, however, the threshold is data dependent and can even degenerate to zero or to one for some data specifications. The result concerning the long run behaviour is obtained for a substantially simplified version of the general model, but the conclusions are completely supported by numerous numerical tests carried out for more general specifications.

The exposition is organized as follows. In Section 2 we describe a general SIS model in its homogeneous and heterogeneous (distributed) versions, and discuss the assumptions. In Section 3 we investigate the dynamics of the integral moments of the susceptible and of the infected heterogeneous populations. Then in Section 4 we pass from the heterogeneous model to a homogeneous one with a nonlinear "prevalence-to-incidense" function and support by analytical and numerical arguments the relevance of the proposed class of approximate incidence functions. Section 5 is devoted to the study of the effect of the heterogeneity on the evolution of the disease in the sort and in the long run.

2 The Homogeneous and the Heterogeneous Models

The model below is of SIS-type, that is, it involves only susceptible and infected individuals. This is the core component of many more detailed epidemiological models being at the same time structurally simplest. On the other hand, within this structure, the model is a rather general one, since nonlinear dependence of fertility, mortality and recovery rates on the population size is allowed.

2.1 The homogeneous model

The main variables in the model are the size $S(t)$ of the susceptible population, and the size $I(t)$ of the infected population at time t.

The dynamics of the disease is described by the equations

$$\dot{S}(t) = -\iota p y S + \lambda(S,I)S + \gamma(S,I)I, \qquad S(0) = S_0, \tag{1}$$

$$\dot{I}(t) = \iota p y S - \delta(S,I)I, \qquad I(0) = I_0, \tag{2}$$

where

$$y(t) = \frac{I(t)}{S(t) + I(t)}.$$

Here
$\lambda = \beta - \mu$ is the net inflow rate, where
β is the birth rate, μ is the mortality rate of the susceptible individuals;
$\delta = \nu + \gamma$ is the out-flow rate of infected individuals[3], where
ν is the mortality rate of the infected individuals, γ is the recovery rate from the infection.

The rate of infection $\iota p y(t)$ consists of three multipliers:
ι is the strength of infection,
p is the (average) level of risk of the population, depending on the average intensity of participation in risky interactions, on the average immunity, etc.,
$y(t)$ is the prevalence of the disease at t.

There is a weak point in the models that include the above one as a kernel component, especially in the expansion phase of the disease and if the population is considerably heterogeneous, in which case the individual values of p may be rather different from the average. In reality, individuals who are more vulnerable to risk become infected with higher probability than average. For many diseases individuals who are more vulnerable as susceptible are more infective if they become infected. As a result, the rate of infection in the early stage of the disease tends to be higher than $\iota p y(t)$. On the other hand, if the mortality of the infected individuals is higher than that of the susceptible ones and the recovered individuals do not increase their level of risk after recovery, then the average vulnerability to risk in the susceptible population decreases with the time. As a result, the value $\iota p y(t)$ overestimates the rate of infection in the late stage of the expansion phase. In our subsequent analysis we formally support these observations and try to avoid the resulting distortion in predicting the evolution of the disease.

2.2 Modeling the heterogeneity

In this subsection we explicitly take into account the heterogeneity of the population, supposing that the value of p is specific for each individual. Similarly as in Diekmann at al. [4] we introduce a variable ω that characterizes individual features that are relevant to the

[3] We assume that only the susceptible individuals are fertile and that all the newborn individuals are susceptible.

disease. As in [4] the variable ω will be called h-state (h stays for heterogeneity). We assume that $\omega \in \Omega$, where Ω is measurable subset of a finite dimensional space.

Correspondingly we introduce the following variables on Ω:
$\bar{S}(t, \cdot)$ – the density of the susceptible population at time t;
$\bar{I}(t, \cdot)$ – the density of the infected population at t.

Moreover, $p(\omega) \geq 0$ will denote the level of risk at h-state ω.

To avoid confusion we note that $\bar{S}(t, \cdot)$, $\bar{I}(t, \cdot) : \Omega \mapsto \mathbf{R}$ are not probability densities, since their integrals over Ω, denoted further by $S(t)$ and $I(t)$, give the total size of the susceptible and of the infected populations at t.

The dynamic of the heterogeneous population is described by the following model where "dot" means derivation with respect to t:

$$\begin{aligned}
\dot{\bar{S}}(t,\omega) &= -\iota p(\omega) z(t) \bar{S}(t,\omega) - \mu \bar{S}(t,\omega) + \beta \int_\Omega \psi_0(S(t), I(t), \bar{S}(t,\omega), \omega, \omega') \bar{S}(t,\omega') \,\mathrm{d}\omega' \\
&\qquad + \gamma \int_\Omega \psi(S(t), I(t), \bar{S}(t,\omega), \omega, \omega') \bar{I}(t,\omega') \,\mathrm{d}\omega', \\
\dot{\bar{I}}(t,\omega) &= \iota p(\omega) z(t) \bar{S}(t,\omega) - \nu \bar{I}(t,\omega) - \gamma \bar{I}(t,\omega).
\end{aligned}$$

The meaning of the rates μ, ν and γ is as in the previous subsection. It is supposed that these rates are independent of ω. The density $\psi_0(\bar{S}(t,\omega), \omega, \omega')$ represents the "probability" that an offspring of an individual of h-state ω' is of h-state ω. Similarly, $\psi(\bar{S}(t,\omega), \omega, \omega')$ represents the "probability" that an infected individual of h-state ω' passes to an h-state ω after recovery. These probabilities are allowed to depend on the current size of the susceptible population of h-state ω. As before the rates μ, β, ν and γ, but also the densities ψ_0 and ψ may depend on the total susceptible and infected populations $S(t)$ and $I(t)$, which is not explicitly indicated in the above formulae.

In the context of this paper ω has more behavioral than purely biological meaning. That is, it represents habits or vulnerability to risk, rather than natural immunity or frailty. Examples are the HIV disease in Central and South Africa (see e.g. Thieme and Castillo-Chavez [9], Sanderson [8], Feichtinger at al. [5]) and the Hepatitis in communities of heroin users (cf. Gavrila at al. [6]). Therefore we assume that the newborn individuals have the same h-distribution as the current susceptible population. Similar assumption we make also for the recovered individuals. The latter is certainly fulfilled if there is no recovery from the disease, as in our main concerns – AIDS and Hepatitis C.

In other words we assume further that (suppressing the time variable in the notations)

$$\psi_0(S, I, \hat{S}(\omega), \omega, \omega') = \psi(S, I, \bar{S}(\omega), \omega, \omega') = \frac{\bar{S}(\omega)}{S}. \tag{3}$$

The term $z(t)$ represents the infectivity of the environment in which the susceptible individuals live. It is called *weighted prevalence* and is defined as

$$z(t) = \frac{J(t)}{R(t) + J(t)},$$

where

$$R(t) = \int_\Omega p(\omega)\bar{S}(t,\omega)\,\mathrm{d}\omega, \quad J(t) = \int_\Omega q(\omega)\bar{I}(t,\omega)\,\mathrm{d}\omega.$$

That is, $z(t)$ is the probability to randomly pick up an infected individual out of the pool of all individuals, if the individuals are counted according to their weights $p(\omega)$, for the susceptible, and $q(\omega)$, for the infected individuals. In the sequel we assume the so-called *proportioned mixing*, Barbour [1]:

$$q(\omega) = \kappa p(\omega), \tag{4}$$

which seems reasonable in the present context[4].

The overall heterogeneous model, under the simplifying assumptions (3),(4), and with the notations λ and δ from the previous section, becomes

$$\begin{aligned}
\dot{\bar{S}}(t,\omega) &= -\iota p(\omega)\frac{J(t)}{R(t)+J(t)}\bar{S}(t,\omega) + \lambda(S,I)\bar{S}(t,\omega) + \gamma(S,I)\frac{I}{S}\bar{S}(t,\omega), & (5)\\
\dot{\bar{I}}(t,\omega) &= \iota p(\omega)\frac{J(t)}{R(t)+J(t)}\bar{S}(t,\omega) - \delta(S,I)\bar{I}(t,\omega), & (6)\\
S(t) &= \int_\Omega \bar{S}(t,\omega)\,\mathrm{d}\omega, & (7)\\
I(t) &= \int_\Omega \bar{I}(t,\omega)\,\mathrm{d}\omega, & (8)\\
R(t) &= \int_\Omega p(\omega)\bar{S}(t,\omega)\,\mathrm{d}\omega, & (9)\\
J(t) &= \kappa\int_\Omega p(\omega)\bar{I}(t,\omega)\,\mathrm{d}\omega, & (10)
\end{aligned}$$

with initial conditions

$$\begin{aligned}
\bar{S}(0,\omega) &= \varphi_0^S(\omega)S_0,\\
\bar{I}(0,\omega) &= \varphi_0^I(\omega)I_0.
\end{aligned}$$

The initial conditions are given in terms of the initial size of the susceptible and of the infected sub-populations, S_0 and I_0, respectively, and the probabilistic densities $\varphi_0^S(\cdot)$ and $\varphi_0^I(\cdot)$ of their h-distributions.

The measurable mapping $(\bar{S}(\cdot,\cdot), \bar{I}(\cdot,\cdot), S(\cdot), I(\cdot), R(\cdot), J(\cdot))$ is a solution of (5)–(10) if for almost every ω the functions $\bar{S}(\cdot,\omega)$ and $\bar{I}(\cdot,\omega)$ are absolutely continuous and (5)–(10) hold

[4] Assumptions (3),(4) are easily justifiable, for example, if the level of risk is determined by the frequency with which the individual is involved in a risky interaction, and if this frequency does not change (or changes proportionally) when the individual becomes infected.

almost everywhere. Below we suppose that the functions λ, γ and δ are at least continuous, or several times continuously differentiable, wherever appropriate. Moreover, we assume that Ω is a closed subset of $\mathbf{R}^r$ with positive Lebesgue measure, that p is a nonnegative measurable function on Ω, and that

$$\mathrm{meas}\{(\omega_1, \omega_2) \in \Omega \times \Omega : \; p(\omega_1) = p(\omega_2)\} = 0. \tag{11}$$

This condition is obviously fulfilled if $\Omega = [0, 1]$ and p is strictly monotone, or if $\Omega = [0, 1]^r$ and p is continuous and strictly monotone along every line through the origin.

Remark. A homogeneous population could be considered as a particular case where $\varphi_0^S(\cdot)$ and $\varphi_0^I(\cdot)$ are concentrated at a single point ω, that is $\varphi_0^S(\cdot) = \varphi_0^I(\cdot) = \delta_\omega(\cdot)$, where $\delta_\omega(\cdot)$ is the Dirac delta function. An alternative way is to take arbitrary φ_0^S and φ_0^I and a constant function $p(\cdot)$. In both cases we come up with a model equivalent to (1),(2), but in the second way involvement of Radon measures is not necessary.

The above heterogeneous model is designed to avoid the shortcomings of the homogeneous models mentioned in the end of the previous subsection. However it has its own disadvantages: (i) it involves distributed parameter integro-differential equations, therefore is more complicated for calculation (especially in its age-structured version, which we do not discuss in this paper); (ii) it requires the initial densities $\varphi_0^S(\cdot)$ and $\varphi_0^I(\cdot)$ which are usually not available.

The first disadvantage is not critical in simulation tasks[5] but could be an obstacle for the design of optimal treatment or prevention strategies, especially in the age-structured case[6]. The second disadvantage makes the direct use of the model impossible even for simulation tasks, due to the lack of data. For the HIV disease in Botswana, for example, even the initial data S_0 and I_0 are vague, while for the densities $\varphi_0^S(\cdot)$ and $\varphi_0^I(\cdot)$ there is no data available at all (cf. Sanderson [8]).

3 Approximating the Heterogeneous System by an Infinite System of ODEs

For the two reasons formulated in the end of the previous sectionwe study the issue of approximation of the heterogeneous model by a homogeneous one, in such a way that the results obtained by the homogeneous model are similar to those that could be obtained by the heterogeneous model if the relevant data were available.

[5] The author has developed an efficient solver for rather general (descriptive or control) heterogeneous models, see Veliov [10].

[6] Which should be the case in modeling HIV in the African countries.

Integrating (5) and (6) with respect to ω we obtain

$$\begin{aligned} \dot{S} &= -\iota\rho^*(t)S + \lambda(S,I)S + \gamma(S,I)I, \quad & S(0) = S_0, \qquad (12)\\ \dot{I} &= \iota\rho^*(t)S - \delta(S,I)I, & I(0) = I_0, \qquad (13) \end{aligned}$$

where

$$\rho^*(t) = z(t)\frac{R(t)}{S(t)} = \frac{J(t)}{R(t)+J(t)}\frac{R(t)}{S(t)}. \qquad (14)$$

Note that the only information that one needs to recover the aggregated solutions $S(\cdot)$ and $I(\cdot)$ of the heterogeneous system by solving the ODE system (12),(13) is the function $\rho^*(t)$.

Define the normalized moments

$$m_k^S(t) = \int_\Omega (p(\omega))^k \frac{\bar{S}(t,\omega)}{S(t)}\,\mathrm{d}\omega, \quad m_k^I(t) = \int_\Omega (p(\omega))^k \frac{\bar{I}(t,\omega)}{I(t)}\,\mathrm{d}\omega, \quad k = 0, 1, \ldots .$$

Obviously

$$R(t) = m_1^S(t)S(t), \quad J(t) = \kappa m_1^I(t)I(t), \qquad (15)$$

From here we obtain

$$z(t) = \frac{\kappa m_1^I(t)I(t)}{m_1^S(t)S(t) + \kappa m_1^I(t)I(t)}, \qquad (16)$$

$$\rho^*(t) = z(t)m_1^S(t) = \frac{\kappa m_1^S(t)m_1^I(t)y(t)}{(1-y(t))m_1^S(t) + \kappa y(t)m_1^I(t)}, \qquad (17)$$

where, as in Section 2.1, we use the notation $y(t) = I(t)/(S(t)+I(t))$ for the prevalence.

Using equations (5) and (12) in the expression for $\dot{m}_k^S$ one can obtain the following equations:

$$\begin{aligned} \dot{m}_1^S &= -\iota z(t)(m_2^S - m_1^S m_1^S),\\ \dot{m}_2^S &= -\iota z(t)(m_3^S - m_2^S m_1^S),\\ \ldots & \quad \ldots\ldots\ldots\ldots\ldots\ldots\\ \dot{m}_k^S &= -\iota z(t)(m_{k+1}^S - m_k^S m_1^S),\\ \ldots & \quad \ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

and similarly, using (6) and (13),

$$\begin{aligned} \dot{m}_1^I &= \iota z(t)\frac{1-y(t)}{y(t)}(m_2^S - m_1^I m_1^S),\\ \dot{m}_2^I &= \iota z(t)\frac{1-y(t)}{y(t)}(m_3^S - m_2^I m_1^S),\\ \ldots & \quad \ldots\ldots\ldots\ldots\ldots\ldots\\ \dot{m}_k^I &= \iota z(t)\frac{1-y(t)}{y(t)}(m_{k+1}^S - m_k^I m_1^S),\\ \ldots & \quad \ldots\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

The initial conditions

$$m_k^S(0) = \int_\Omega (p(\omega))^k \varphi_0^S(\omega)\,\mathrm{d}\omega, \quad m_k^I(0) = \int_\Omega (p(\omega))^k \varphi_0^I(\omega)\,\mathrm{d}\omega$$

are known, provided that the initial distributions φ_0^S and φ_0^I are given.

Having in mind (16) and (17) we establish that the above infinite system of differential equations, together with (12),(13) determines the solution $(S(\cdot), I(\cdot))$. It can be used for numerical approximation of $(S(\cdot), I(\cdot))$ in a version of the method of Poincaré by a truncation of the infinite system. Such an approximating procedure, however, still makes use of the initial data, $\varphi_0^S(\cdot)$ and $\varphi_0^I(\cdot)$ which should be avoided, as we argued in the end of Section 2, therefore we do not discuss the details.

Another consequence of the above reformulation of the heterogeneous system is the following.

Proposition 1 *Assume that $J(0) > 0$. Then for each $k \geq 1$ the moment $m_k^S(\cdot)$ is strictly monotone decreasing. For $k > 1$ also the normalized moment $m_k^S(\cdot)/m_1^S(\cdot)$ is strictly monotone decreasing. Moreover, if $m_k^S(t) \leq m_k^I(t)$ for $t = 0$, then this inequality holds for all $t \geq 0$.*

Proof. The proof makes use of the following Lyapunov type inequalities: if $\phi : \Omega \mapsto \mathbf{R}$ is nonnegative, measurable probability distribution, and

$$m_k \overset{def}{=} \int_\Omega (p(\omega))^k \phi(\omega)\,\mathrm{d}\omega,$$

then for every nonnegative integers $p < q < r < s$ with $p + s = q + r$

$$m_p m_s > m_q m_r. \tag{18}$$

Here the inequality is strict due to assumption (11). Applying this inequality with $p = 0$, $s = k + 1$, $q = k$, $r = 1$ we obtain $m_{k+1}^S - m_k^S m_1^S > 0$. From $J(0) > 0$ and (6) it follows that $J(t) > 0$ for every t. Thus $z(t) > 0$, which implies $\dot{m}_k^S(t) < 0$.

To prove the second claim we calculate

$$\left(\frac{m_k^S}{m_i^S}\right)' = -\iota z \frac{(m_{k+1}^S - m_k^S m_1^S)m_i^S - (m_{i+1}^S - m_i^S m_1^S)m_k^S}{(m_i^S)^2} = -\iota z \frac{m_{k+1}^S m_i^S - m_{i+1}^S m_k^S}{(m_i^S)^2}.$$

The last quantity is strictly negative according to (18) applied with $p = i$, $q = i + 1$, $r = k$, $s = k + 1$.

To prove the invariance of the area $m_k^S \leq m_k^I$ with respect to the differential equations for the moments it is enough (this is obvious, however, for a strict reasoning one can apply the Nagumo invariance theorem) to verify that $\dot{m}_k^S(t) \leq \dot{m}_k^I(t)$ whenever $m_k^S(t) = m_k^I(t)$. At such a point t we have $\dot{m}_k^S(t) \leq 0$ and $\dot{m}_k^I(t) \geq 0$ by the same inequality used in the proof of the first claim. Q.E.D.

Corollary 1 *The heterogeneous system (5)–(10) does not have a periodic solution with $J(0) > 0$, whatever are the functions λ, γ and δ.*

We mention that the homogeneous system (1),(2) may have a periodic solution for appropriate λ, γ and δ depending on S and I.

The value

$$\frac{m_2^S(t)}{m_1^S(t)} - m_1^S(t)$$

(which is "variance/mean" of $p(\cdot)$) can be viewed as a measure of heterogeneity of the current susceptible population. If the two populations have the same mean risk $m_1^S(0)$ at time 0, then the one with the higher value of $m_2^S(0)$ is more heterogeneous.

4 Encapsulating Heterogeneity in a Homogeneous Model

In this section we continue the analysis of the heterogeneous model (5)–(10) aiming at obtaining a homogeneous system of the form of (1),(2) which simulates the heterogeneous one. The key point is that we replace the multiplier y in (1),(2) with a nonlinear function of the prevalence, $\rho(y)$. It turns out that there exists such a function $\rho(y)$ for which the solution of (1),(2) (with y replaced by $\rho(y)$) coincides with the (S, I)-part of the solution of the heterogeneous system (5)–(10). The advantage of knowing the appropriate function $\rho(y)$ is that one would not need to know the distributions φ_0^S and φ_0^I in order to simulate the evolution of the disease in a heterogeneous population. Our approach suggests to approximate the function $\rho(y)$ by measuring the prevalence $y(t)$ at several moments t and applying standard identification technique. For this purpose one has to restrict the search of the function $\rho(y)$ to a class of functions Γ depending on a few parameters. In order to justify a choice of the class Γ we first establish some qualitative properties of the function $\rho(y)$.

We stress that the approach below is appropriate for the expansion phase of the disease, where the prevalence $y(t)$ is increasing. Let $(\bar{S}(\cdot,\cdot), \bar{I}(\cdot,\cdot), S(\cdot), I(\cdot), R(\cdot), J(\cdot))$ be a solution of the heterogeneous system of (5)–(10), let $\rho^*(t)$ be the function defined in (14), and $y(t) = I(t)/(S(t) + I(t))$ be the prevalence. We suppose that $\dot{y}(0) > 0$, thus there is an interval $[0,t^*)$ (t^* could be $+\infty$) such that $\dot{y}(t) > 0$ on $[0,t^*)$ and $\dot{y}(t^*) = 0$ (in the case $t^* = +\infty$ the last equation should be disregarded). Let $[y_0, y_*)$ be the set of values of $y(t)$ when t runs in $[0,t^*)$.

Thanks to the strict monotonicity of $y(\cdot)$, the equation

$$\rho(y(t)) = \rho^*(t) \tag{19}$$

defines in a unique way the function $\rho : [y_0, y_*) \mapsto \mathbf{R}$. Then the solution of the system

$$\dot{S} = -\iota\rho\left(\frac{I}{S+I}\right)S + \lambda(S,I)S + \gamma(S,I)I, \qquad S(0) = S_0, \tag{20}$$

$$\dot{I} = \iota\rho\left(\frac{I}{S+I}\right)S - \delta(S,I)I, \qquad I(0) = I_0, \tag{21}$$

coincides with the (S,I)-part of the solution of the heterogeneous model (5)–(10). The definition of the nonlinear "prevalence-to-incidence" function ρ, however, is not constructive, since it requires knowledge of the solution of (5)–(10). Therefore the next step to constructive approximation will be to establish some properties of the function $\rho(y)$. In doing this we make the following simplifying assumption:

$$m_1^S(0) \geq m_1^I(0). \tag{22}$$

(If this is an equality, it would mean that the disease is initiated in a random way.)

Solving the differential equation for m_1^I and denoting $q(t) = \iota z(t)(1-y(t))/y(t)$ we have

$$m_1^I(t) = m_1^I(0)e^{-\int_0^t q(\theta)m_1^S(\theta)d\theta} + \int_0^t e^{-\int_\xi^t q(\theta)m_1^S(\theta)d\theta} q(\xi)m_2^S(\xi)\,\mathrm{d}\xi$$

and integrating by parts and rearranging the terms we obtain

$$m_1^I(t) - \frac{m_2^S(t)}{m_1^S(t)} = \left(m_1^I(0) - \frac{m_2^S(0)}{m_1^S(0)}\right)e^{-\int_0^t q(\theta)m_1^S(\theta)d\theta} - \int_0^t \frac{\mathrm{d}}{\mathrm{d}\xi}\left(\frac{m_2^S}{m_1^S}\right)(\xi)e^{-\int_\xi^t q(\theta)m_1^S(\theta)d\theta}\,\mathrm{d}\xi$$

Substituting this in the equation for m_1^I we have

$$\begin{aligned}&\dot{m}_1^I(t)\\ &= q(t)m_1^S(t)e^{-\int_0^t q(\theta)m_1^S(\theta)d\theta}\left[\left(\frac{m_2^S(0)}{m_1^S(0)} - m_1^I(0)\right) + \int_0^t \frac{\mathrm{d}}{\mathrm{d}\xi}\left(\frac{m_2^S}{m_1^S}\right)(\xi)e^{\int_0^\xi q(\theta)m_1^S(\theta)d\theta}\,\mathrm{d}\xi\right].\end{aligned}$$

The first term in the last brackets is constant and positive. The derivative of m_2^S/m_1^S is strictly negative according to Proposition 1. Therefore, the second term is zero at $t = 0$ and monotone decreasing. We come to the following.

Claim 1: If $y(0) \in (0,1)$ and $t^* > 0$, then there exists $t_I \in [0,t^*]$ such that $\dot{m}_1^I(t) > 0$ on $[0,t_I)$ and $\dot{m}_1^I(t) < 0$ on (t_I, t^*).

It may happen that $t_I = t^*$, that is, m_1^I is increasing in the whole expansion phase.

From the definition of the function ρ we have

$$\rho'(y(t)) = \frac{\dot{\rho}^*(t)}{\dot{y}(t)}.$$

Having that (see (17))

$$\rho^*(t) = \frac{\kappa m_1^S m_1^I y}{(1-y)m_1^S + \kappa y m_1^I}(t),$$

one obtains

$$\rho'(y(t)) = \kappa \frac{m_1^I (m_1^S)^2 + \dot{m}_1^I (m_1^S)^2 y(1-y)/\dot{y} + \kappa \dot{m}_1^S (m_1^I)^2 y^2/\dot{y}}{\left[(1-y)m_1^S + \kappa y m_1^I\right]^2}(t).$$

Claim 2: If $y(0) = y_0 \in (0, (1+\sqrt{\kappa})^{-1})$, then the function $\rho(\cdot) : [y_0, y^*) \mapsto \mathbf{R}$ is strictly increasing close to $y = y_0$. If $t^* < +\infty$ and either $y^* = 1$ or $t_I < t^*$, then $\rho(\cdot)$ is strictly decreasing close to y^*.

To prove the first part of this claim we substitute $\dot{m}_1^S$ and m_1^I by the right-hand sides of the corresponding differential equations in Section 3. This gives

$$\rho'(y(t)) = \kappa \frac{m_1^I (m_1^S)^2 + \frac{\iota z}{\dot{y}} \left[(1-y)^2 (m_2^S - m_1^I m_1^S)(m_1^S)^2 - \kappa y^2 (m_2^S - (m_1^S)^2)(m_1^I)^2\right]}{\left[(1-y)m_1^S + \kappa y m_1^I\right]^2}$$

Using the inequality (22) we obtain for $t = 0$

$$\rho'(y_0) \geq \kappa \frac{m_1^I (m_1^S)^2 + \frac{\iota z}{\dot{y}} [m_2^S - (m_1^S)^2](m_1^I)^2 \left[(1-y)^2 - \kappa y^2\right]}{\left[(1-y)m_1^S + \kappa y m_1^I\right]^2},$$

which implies the first statement in Claim 2.

To prove the second part we notice that $\dot{\bar{y}}(t^*) = 0$. Moreover, if $t_I < t^*$, then $m_1^I(t) < 0$ close to t^*. Then the sign of ρ' close to y^* is determined by that of $\dot{m}_1^I (m_1^S)^2 y(1-y) + \dot{m}_1^S (m_1^I)^2 y^2$, evaluated at $t = t^*$. The second term is negative, according to of Proposition 1The first term is either zero (if $y^* = 1$) or negative (if $t_I < t^*$), according to Claim 1.

The above consideration implies also the following.

Claim 3: $\rho(\cdot)$ has a bounded derivative in every compact subinterval of $[y_0, y^*)$, but $\rho'(y)$ may converge to $-\infty$ at y^*.

We mention that it is a known fact that a nonlinear “prevalence-to-incidense” function ρ may be more relevant (due to the heterogeneity) than a linear one, like in (20)–(21). Sanderson

[8], for example uses the function $\rho(y) = ay^{\alpha}$ in the context of the HIV disease in Botswana, with α identified from real data as $\alpha \approx 0.3$. Notice, however, that this function does not fulfill neither Claim 2 nor Claim 3. Indeed, it is everywhere monotone increasing and has infinite derivative at $y = 0$.

Based on the above properties and on a number of numerical experiments with the heterogeneous model we propose the following class of functions depending on four parameters a, α, b, β to be used as an approximation of the function $y \longrightarrow \rho(y)$:

$$\rho(y) \overset{def}{=} ay^{\alpha}(1 - by)^{\beta}, \tag{23}$$

where

$$a \geq 0, \quad \alpha \in [0, 1], \quad b \geq 1, \quad \beta \in [0, 1].$$

Here, in fact, $b = 1/y^*$. Notice that according to Claim 3 we may fix $\alpha = 1$. We deliberately keep a redundancy by allowing values of $\alpha < 1$, but we shall see later that identifying the four parameters from experimental data leads to $\alpha = 1$ (or to a value close to one, in some exceptional cases). Thus, in fact, only the three parameters a, b, β need to be identified from data in a real application. Notice that no data about the heterogeneity of the population is required. Provided that the other parameters in (20), (21) are known, one needs only measurements of $y(t)$ for at least three moments of time.

We illustrate the proposed approximation by several examples, where simulation of the heterogeneous system (5)–(10) is used for determination of the "true" "prevalence-to-incidense" function $\rho(y)$ by (14), (19).

Figure 1 shows the "true" function $\rho(y)$ arising from a heterogeneous model with artificial fertility/mortality data (roughly calibrated for a human population). In all plots the mean risk at time 0, $(m_1^S(0), m_1^I(0))$, is the same. The only difference is the level of heterogeneity $m_2^S(0) = m_2^I(0)$. Bigger values of h correspond to higher heterogeneity. The value $h = 0$ corresponds to a homogeneous population, and the function $\rho(y)$ is linear in this case. The higher is the heterogeneity, the more significant is the deviation from the straight line. Notice that for $h = 0.8$ still $y^* = 1$, but for $h = 1$ we see $y^* < 0.9$.

[Figure 1 about here.]

Figure 2 represents the functions $\rho(y)$ for two different levels of heterogeneity, together with their best uniform approximations (the dotted lines) with functions from the class (23)[7]. Here measurements of $\rho(y)$ are taken over the whole interval $[y_0, y^*)$. In figures 3 and 4, only five (resp. ten) measurement in the intervals indicated by vertical lines are used for

[7] One can clearly see two Chebishev points – a fact that, to our knowledge does not follow from a known theoretical results, but can be intuitively expected in view of the chosen class of functions (23).

finding the parameters in (23). The dotted line represents the best uniform approximation in the class of functions (23). The dash-dotted lines represent the best approximation based on the same data, in the class of functions used in Sanderson (2002). The forecast based on the dash-dotted extrapolation of $\rho(\cdot)$ would be rather pessimistic. Notice also that in Figure 4 the function $\rho(\cdot)$ is rather flat in the interval $[0.1, 0.2]$ where the measurements are taken. Nevertheless, the tendency of future decrease is captured by the approximating class (23).

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

5 The Effect of the Heterogeneity on the Evolution of an Epidemic Disease

Now we shall investigate how the level of heterogeneity of the population influences, *ceteris paribus*, the evolution of the disease. It turns out that the effect of the heterogeneity on a short and on a long time horizons could be opposite, therefore we study these two cases separately.

5.1 Short run

Proposition 2 *Let $(\bar{S}, \bar{I}, S, I, R, J)$ and $(\tilde{\bar{S}}, \tilde{\bar{I}}, \tilde{S}, \tilde{I}, \tilde{R}, \tilde{J})$ be two solutions of (5)–(10) corresponding to initial data $(S_0\varphi_0^S(\cdot), I_0\varphi_0^I(\cdot))$ and $(S_0\tilde{\varphi}_0^S(\cdot), I_0\tilde{\varphi}_0^I(\cdot))$, such that $m_1^S(0) = \tilde{m}_1^S(0)$, $m_1^I(0) = \tilde{m}_1^I(0)$ and $m_2^S(0) < \tilde{m}_2^S(0)$. If*

$$\frac{J(0)}{R(0) + J(0)} < \frac{1}{2} \tag{24}$$

then there exists $\bar{t} > 0$ such that

$$S(t) > \tilde{S}(t), \quad I(t) < \tilde{I}(t), \quad \forall\, t \in (0, \bar{t}]. \tag{25}$$

If the opposite strict inequality holds in (24), then there exists $\bar{t} > 0$ such that the opposite strict inequalities hold in (25).

Proof. Let us denote $\Delta_S(t) = \tilde{S}(t) - S(t)$ and $\Delta_I(t) = \tilde{I}(t) - I(t)$. Similar notation $\Delta_g(t)$ will be used also for other variables g.

Obviously $\tilde{\rho}^*(0) = \rho^*(0)$, therefore from the conditions of the proposition and (12),(13) we obtain that $\Delta'_S(0) = \Delta'_I(0) = 0$. We shall investigate the sign of $\Delta''_S(0)$. Differentiating (12) and using the above equalities and the assumptions we obtain that

$$\Delta''_S(0) = -(\dot{\tilde{\rho}}^*(0) - \dot{\rho}^*(0))S_0.$$

We have

$$\begin{aligned}\dot{\rho}^*(0) &= \left(z(t)\frac{R(t)}{I(t)}\right)'_{t=0} = z'(0)m_1^S(0) + z(0)\dot{m}_1^S(0) \\ &= z'(0)m_1^S(0) - z^2(0)(m_2^S(0) - m_1^S(0)m_1^S(0)).\end{aligned}$$

The derivative of z can be found by differentiating $J(t)/(R(t)+J(t))$, where the derivatives of R and J are calculated from (9),(10) using also (7),(8). After certain calculations one can represent

$$\begin{aligned}\dot{z}(0) &= m_2^S(0)\frac{S_0 z(0)}{R(0)+J(0)} - \zeta, \\ \dot{\tilde{z}}(0) &= \tilde{m}_2^S(0)\frac{S_0 \tilde{z}(0)}{\tilde{R}(0)+\tilde{J}(0)} - \zeta,\end{aligned}$$

where ζ is the same in the two formulas. Notice that $\tilde{R}(0) = \tilde{m}_1^S(0)S_0 = m_1^S(0)S_0 = R(0)$ and similarly for J. Also $\tilde{z}(0) = z(0)$. Then we obtain

$$\begin{aligned}\Delta''_S(0) = &\left[m_2^S(0)\frac{S(0)z(0)}{R(0)+J(0)} - \zeta\right]\frac{R(0)}{S(0)} - z^2(0)(m_2^S(0) - m_1^S(0)m_1^S(0)) \\ &- \left[\left[\tilde{m}_2^S(0)\frac{\tilde{S}(0)\tilde{z}(0)}{\tilde{R}(0)+\tilde{J}(0)} - \zeta\right]\frac{\tilde{R}(0)}{\tilde{S}(0)} - \tilde{z}^2(0)(\tilde{m}_2^S(0) - \tilde{m}_1^S(0)\tilde{m}_1^S(0))\right] \\ = &\; z(0)\left[\frac{R(0)}{R(0)+J(0)} - z(0)\right](m_2^S(0) - \tilde{m}_2^S(0)).\end{aligned}$$

Since the last multiplier is negative, $\Delta''_S(0)$ will be negative (therefore $\Delta_S(t)$ will be negative for small t) if the multiplier in the brackets is positive. The latter is equivalent to $R(0) > J(0)$, which coincides with (24). Q.E.D.

The meaning of the above proposition can be explained in the following informal way. Let us compare two populations with the same mortality, birth and recovery rates, with the same initial size and same initial number of infected individuals, and also with the same mean level of risk at time $t = 0$. Let only the initial level of heterogeneity of the susceptible sub-populations are different: $m_2^S(0) < \tilde{m}_2^S(0)$. Then in short run the less heterogeneous population has

(i) more susceptible individuals and less infected individuals, if the weighted prevalence at $t = 0$ is less than 0.5;
(ii) less susceptible individuals and more infected individuals, if the weighted prevalence at $t = 0$ is bigger than 0.5.

The fact that for a very high weighted prevalence, z(0), the higher heterogeneity is advantageous is obvious. Not that obvious is that for very low weighted prevalence the lower heterogeneity is advantageous in a short run. What is, to our opinion, somewhat unexpected, is that the threshold value is always $z = 0.5$, which is independent of the particular functions λ, γ and δ, as well as on the other data involved.

The result is illustrated in figures 5, 6 for initial weighted prevalence $z(0) < 0.5$, and in Fig. 7 for $z(0) > 0.5$.

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

5.2 Long run

From figures 5 and 7 one may expect that the more heterogeneous population will maintain more susceptible individuals in the long run. This conclusion, however, is false. The analysis of the asymptotic behaviour of the solution of (5)–(10) is a complicated task at this level of generality. Therefore we shall consider a simplified situation where we shall obtain a complete analytic description of the asymptotic behaviour, depending on the initial prevalence.

Namely, we compare two populations with the same initial number of infected individuals and the same mean risk at time zero. One of these population is assumed homogeneous, with $p(\omega) = p$ for all individuals. The second population is assumed heterogeneous with the set $\Omega = \{\omega_1, \omega_2\}$ consisting of only two values: at time zero a fraction α of the individuals have no risk at all ($p(\omega_1) = 0$), the rest have risk level $p_2 > 0$. In order to have the same mean value of risk at time $t = 0$ we have to take $p_2 = p/(1 - \alpha)$. Moreover, we assume that there is no recovery, that is, $\gamma = 0$, that the rates λ and δ are constants, and that $\kappa = 1$.

The equations of the homogeneous population become

$$\begin{aligned} \dot{S}(t) &= -\iota p \frac{I}{S+I} S + \lambda S, \qquad S(0) = S_0, \\ \dot{I}(t) &= \iota p \frac{I}{S+I} S - \delta I, \qquad I(0) = I_0. \end{aligned}$$

Denoting $S_i(t) = S(t, \omega_i)$, $I_i(t) = I(t, \omega_i)$, $i = 1, 2$, we rewrite the equations for the heterogeneous population as

$$\begin{aligned}
\dot{S}_1 &= \lambda S_1, \quad S_1(0) = \alpha S_0, \\
\dot{I}_1 &= -\delta I_1, \quad I_1(0) = \alpha I_0, \\
\dot{S}_2 &= -\iota \frac{p}{1-\alpha} \frac{I_2}{S_2 + I_2} S_2 + \lambda S_2, \qquad S_2(0) = (1-\alpha) S_0, \\
\dot{I}_2 &= \iota \frac{p}{1-\alpha} \frac{I_2}{S_2 + I_2} S_2 - \delta I_2, \qquad I_2(0) = (1-\alpha) I_0.
\end{aligned}$$

Each of the above two systems can be transformed to a Ricati-type system and can be solved analytically. The solution depend qualitatively on the quantities

$$r = \lambda + \delta - \iota p, \quad \text{and} \quad r_\alpha = \lambda + \delta - \frac{\iota p}{1-\alpha}.$$

Denoting by $S^{\text{hom}}(t) = S(t)$ and $S^{\text{heter}}(t) = S_1(t) + S_2(t)$ the size of the susceptible part of the homogeneous (respectively heterogeneous) population, and solving the corresponding equations one can obtain

$$S^{\text{hom}}(t) = e^{\lambda t} S_0 \times \begin{cases} \left(y(0)e^{-rt} + 1 - y(0)\right)^{\frac{\iota p}{r}}, & \text{if } r \neq 0, \\ e^{-\iota p y(0) t}, & \text{if } r = 0, \end{cases}$$

$$S^{\text{heter}}(t) = e^{\lambda t} S_0 \times \begin{cases} \alpha + (1-\alpha)\left(y(0)e^{-r_\alpha t} + 1 - y(0)\right)^{\frac{\iota p}{(1-\alpha) r_\alpha}}, & \text{if } r_\alpha \neq 0, \\ \alpha + (1-\alpha) e^{-\frac{\iota p}{1-\alpha} y(0) t}, & \text{if } r_\alpha = 0, \end{cases}$$

As above, $y(0)$ is the initial prevalence in the two populations.

We compare the limits at infinity of the "detrended" populations:

$$S^{\text{hom}}_{\text{det}}(\infty) := \lim_{t \to +\infty} e^{-\lambda t} S^{\text{hom}}(t) = S_0 \times \begin{cases} (1 - y(0))^{\frac{\iota p}{r}}, & \text{if } r > 0 \\ 0, & \text{if } r \leq 0, \end{cases}$$

$$S^{\text{heter}}_{\text{det}}(\infty) := \lim_{t \to +\infty} e^{-\lambda t} S^{\text{heter}}(t) = S_0 \times \begin{cases} \alpha + (1-\alpha)(1 - y(0))^{\frac{\iota p}{(1-\alpha) r_\alpha}}, & \text{if } r_\alpha > 0 \\ \alpha, & \text{if } r_\alpha \leq 0. \end{cases}$$

Three cases could be distinguished:

Case 1. $0 < r_\alpha < r$. By an elementary argument one can prove that the equation

$$(1-y)^{\frac{\iota p}{r}} = \alpha + (1-\alpha)(1-y)^{\frac{\iota p}{(1-\alpha) r_\alpha}}$$

has a unique solution $\hat{y} \in [0, 1]$. Then we have

$$S^{\text{hom}}_{\text{det}}(\infty) > S^{\text{heter}}_{\text{det}}(\infty) \quad \text{if} \quad y(0) < \hat{y}, \tag{26}$$

$$S_{\text{det}}^{\text{hom}}(\infty) < S_{\text{det}}^{\text{heter}}(\infty) \quad \text{if} \quad y(0) > \hat{y}, \tag{27}$$

Case 2. $r_\alpha \leq 0 < r$. In this case the same conclusions (26), (27) hold, with the only difference that $\hat{y}$ is determined from the equation

$$(1-y)^{\frac{\iota p}{r}} = \alpha,$$

which has a unique solution $\hat{y} \in [0, 1]$.

Case 3. $r_\alpha < r \leq 0$. In this case $0 = S_{\text{det}}^{\text{hom}}(\infty) < S_{\text{det}}^{\text{heter}}(\infty)$ holds for every positive initial prevalence $y(0)$.

We summarize the above conclusions as follows: there is a threshold initial prevalence $\hat{y}$, such that the more homogeneous population is asymptotically larger than the more heterogeneous one if $y(0) < \hat{y}$. The converse is true for $y(0) > \hat{y}$. For some values of the parameters the threshold $\hat{y}$ can degenerate to zero (as in Case 3) or to value one.

The above conclusions are drown on the basis of a simplified model, but are also supported by all our experiments with various models of the form (5)–(10).

Thus the dependence of the long run behaviour on the initial (weighted) prevalence is very much similar to that of the short run behaviour: for initial (weighted) prevalence below the threshold, less heterogeneity is advantageous, while above the threshold more heterogeneity is advantageous. The threshold value always equals 0.5 in the sort run consideration, while it is data-dependent in the long run. In both figures 5 and 7, for example, more heterogeneity is advantageous in the long run, which means that for the threshold $\hat{y}$ it holds $\hat{y} \leq 0,98$ and $\hat{y} \leq 0.25$.

References

[1] Barbour, A.D. MacDonald's model and transmission of bilharzia. *Trans. Roy. Soc. Trop. Mad. Hyg.*, **72**:6–15 (1978).

[2] Coutinho, F.A.B., Massad, E., Lopez, L.F., and Burattini, M.N. Modelling Heterogeneity in individual frailties in epidemic models. *Mathematical and Computer Modelling*, **30**:97–115 (1999).

[3] Diekmann, O., and Heesterbeek, J.A.P. *Mathematical epidemiology of infectious diseases. Model building, analysis and interpretation.* Wiley Series in Mathematical and Computational Biology. John Wiley & Sons, Ltd., Chichester, 2000.

[4] Diekmann, O., Heesterbeek, J.A.P., and Metz, J.A.J. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biology*, **28**:365–382 (1990).

[5] Feichtinger, G., Tsachev, Ts., and Veliov, V.M. Maximum principle for age and duration structured systems: a tool for optimal prevention and treatment of HIV. *Mathematical Population Studies* (to appear).

[6] Gavrila, C., Pollack, H.A., Caulkins, J.P., Kort, P.M., Feichtinger, G., and Tragler, G. Optimal control of harm reduction in preventing blood-borne diseases among drug users, In submission.

[7] Murray, J.D. *Mathematical Biology.* Springer, 1989.

[8] Sanderson, W.C. The demographic impact of HIV medication programs: with examples from Botswana. Paper presented at the Population Association of America Meetings, Atlanta, GA, May, 2002.

[9] Thieme H. R., and Castillo-Chavez C. How may infection-age-dependent infectivity affect the dynamics of HIV/AIDS? *SIAM Journal on Applied Mathematics*, **53**:1447–1479 (1993).

[10] Veliov, V.M. Newton's method for problems of optimal control of heterogeneous systems. *Optimization Methods and Software*, **18**(6):689–703 (2003).

List of Figures

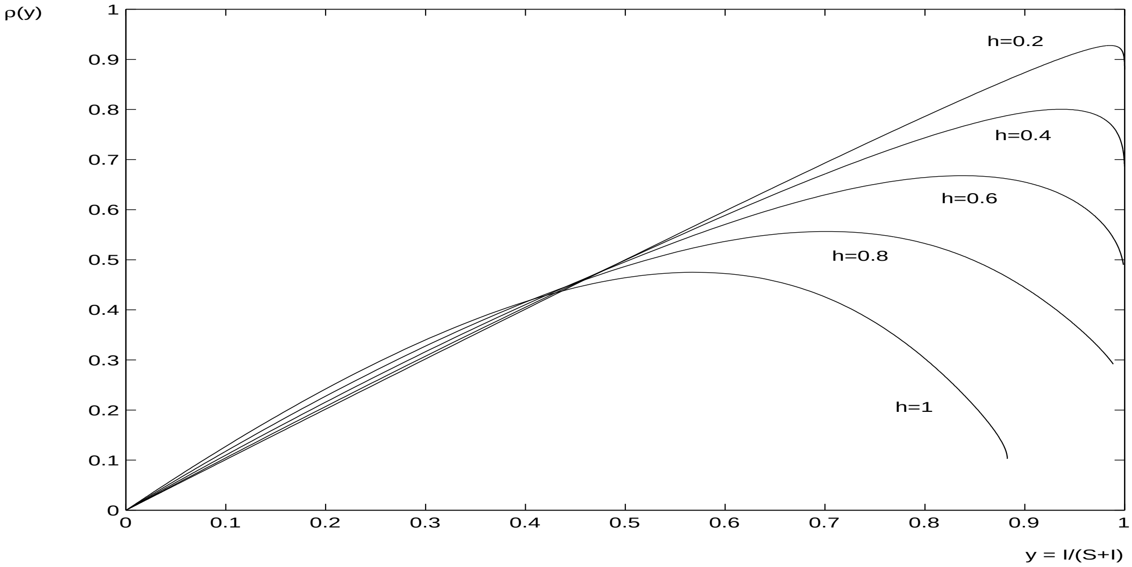

Figure 1: The "prevalence-to-incidense" function $\rho(y)$ for different levels of heterogeneity h: the bigger h, the more heterogeneous the population.

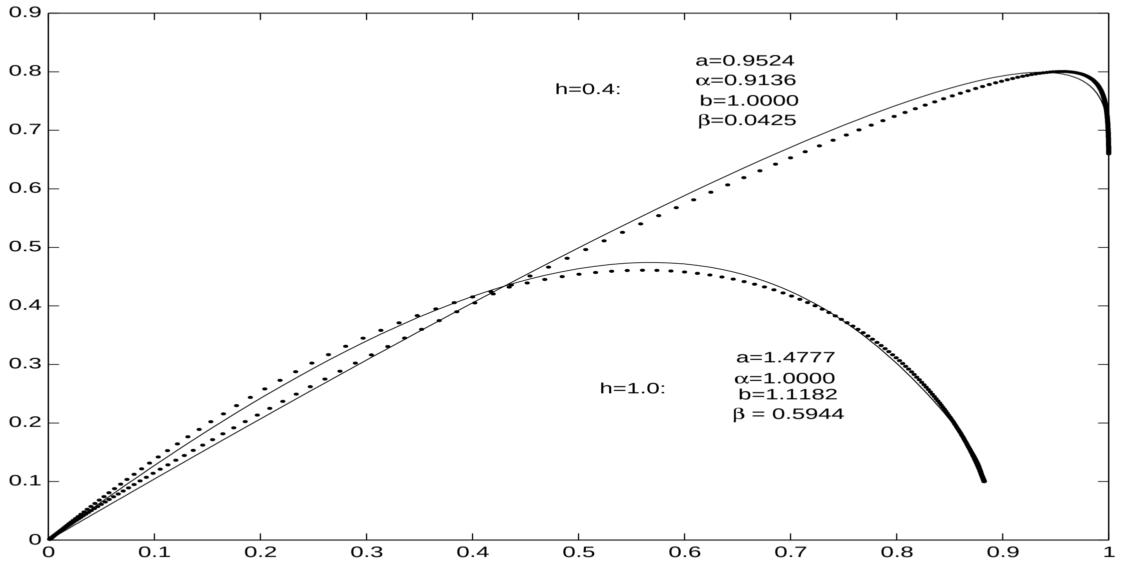

Figure 2: The functions $\rho(\cdot)$ for $h = 1$ and $h = 0.4$ and their best uniform approximations by functions from the class (23) (dotted lines) and by the functions ay^{α} (dash-dotted lines).

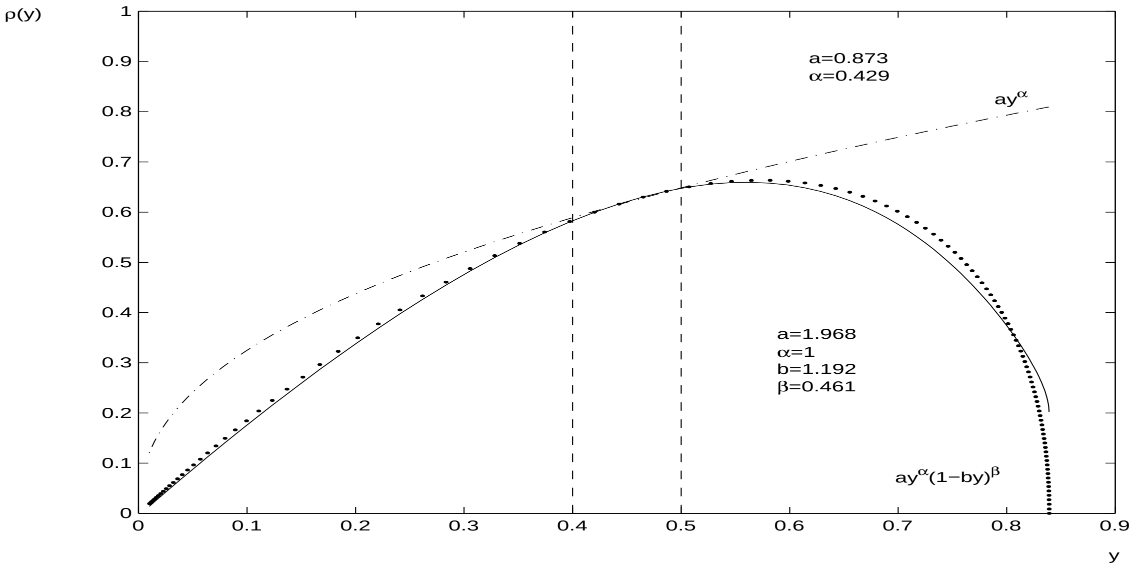

Figure 3: Extrapolations of $\rho(\cdot)$ by functions from the class (23) (dotted line) and by ay^{α} (dash-dotted) from measurements in $[0.4, 0.5]$.

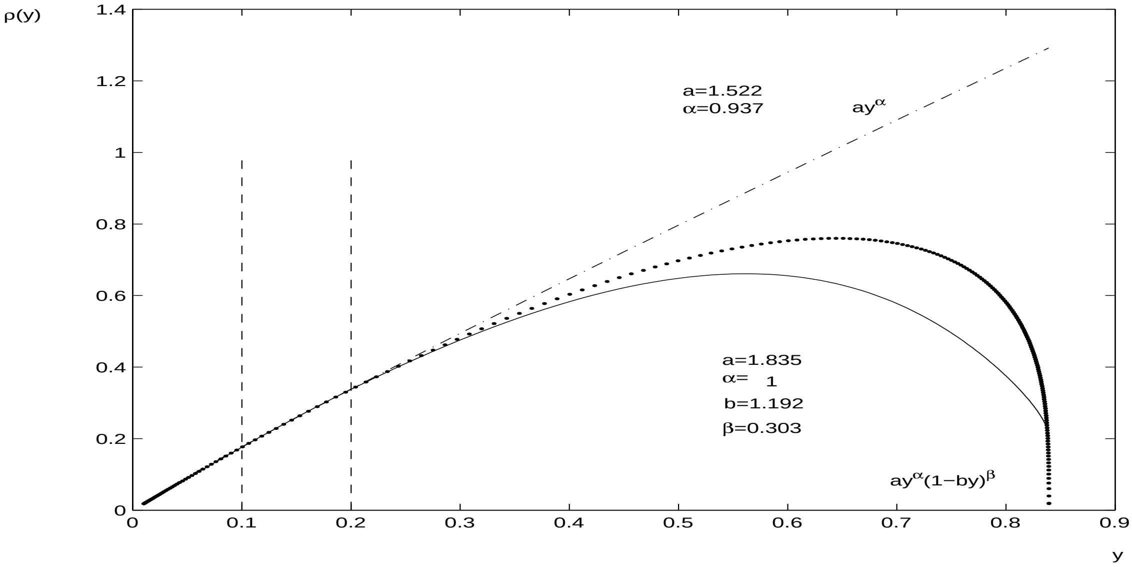

Figure 4: Extrapolations of $\rho(\cdot)$ by functions from the class (23) (dotted line) and by ay^α (dash-doted) from measurements in $[0.1, 0.2]$.

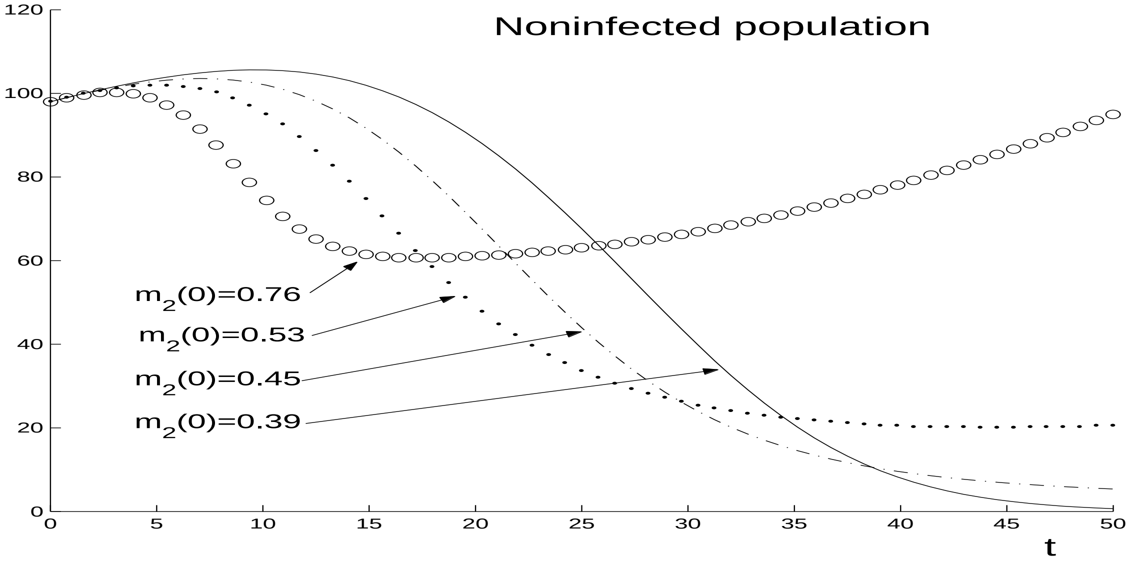

Figure 5: Evolution of the susceptible part of populations differing only with their level of heterogeneity. The initial prevalence is $y(0) = 0.02$.

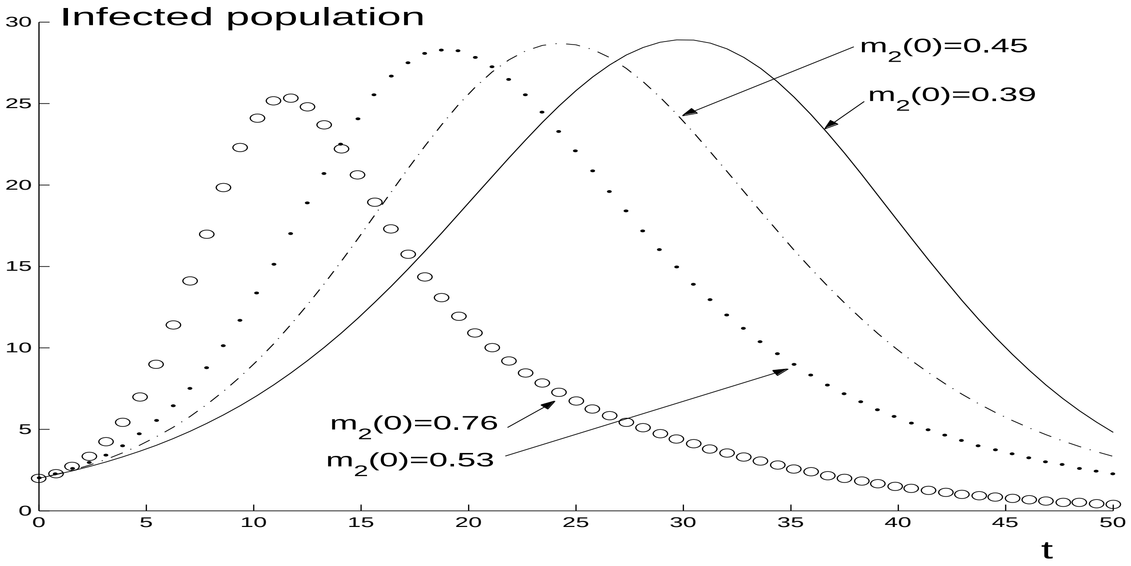

Figure 6: Evolution of the infected part of populations differing only with their level of heterogeneity. The initial prevalence is $y(0) = 0.02$.

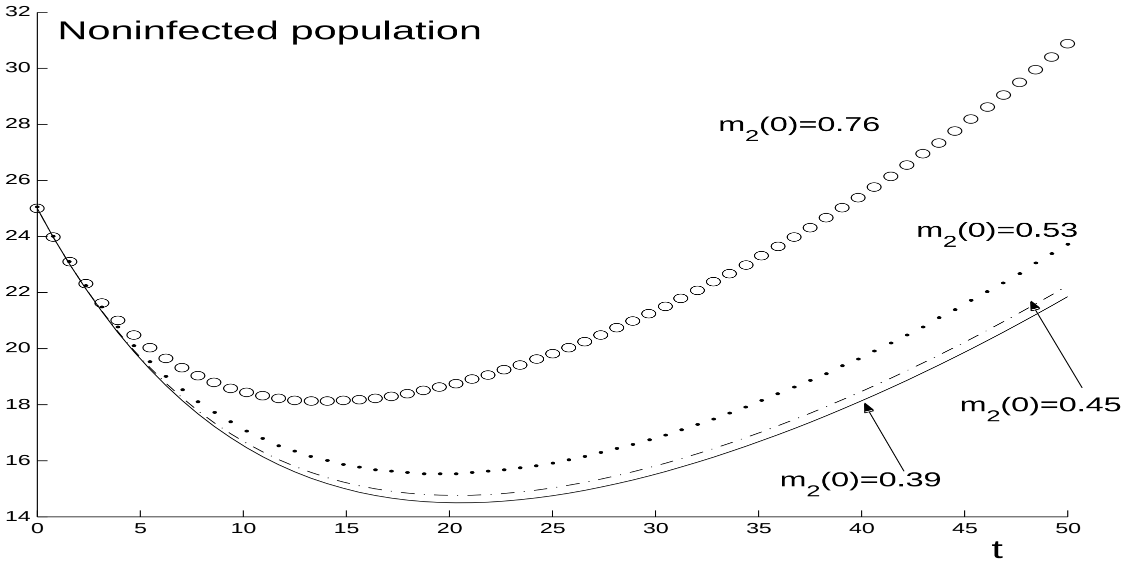

Figure 7: Evolution of the susceptible part of populations differing only with their level of heterogeneity. The initial prevalence is $y(0) = 0.75$.