Title
Cost Optimization in the SIS Model of Infectious Disease with Treatment

Authors
Goldman, Steven M.
Lightwood, James

Publication Date
1996

Peer reviewed

UNIVERSITY OF CALIFORNIA AT BERKELEY

Department of Economics

Berkeley, California 94720-3880

Working Paper No. 96-245

Cost Optimization in the SIS Model of Infectious Disease with Treatment

Steven M. Goldman

and

James Lightwood

January 1996

Key words: Cost optimization, infectious disease, SIS model

Abstract

We consider the intertemporal social optimization problem of minimizing the present value of the costs incurred from both disease and treatment. Though the analysis is complicated by the analytical failure of concavity, we are able to substantially characterize both the long run equilibria and the adjustment paths. The cost minimizing program is shown to exhibit a tendency towards decreased levels of treatment in the presence of higher disease levels. The socially optimal program is compared to individually rational behavior and the inefficiencies in private behavior from the infection externality are shown to cause potentially large increases in the equilibrium rate of infection.

JEL Classification: H01

Cost Optimization in the SIS Model of Infectious Disease with Treatment

Steven M. Goldman James Lightwood

October 31, 1995

Abstract

We consider the intertemporal social optimization problem of minimizing the present value of the costs incurred from both disease and treatment. Though the analysis is complicated by the analytical failure of concavity , we are able to substantially characterize both the long run equilibria and the adjustment paths. The cost minimizing program is shown to exhibit a tendancy towards decreased levels of treatment in the presence of higher disease levels. The socially optimal program is compared to individually rational behavior and the inefficiencies in private behavior from the infection externality are shown to cause potentially large increases in the equilibrium rate of infection.

1. Introduction to Economics Of Medical Treatment In The SIS Infectious Disease Model

1.1. Background

The control of infectious disease has been one of the most dramatic successes of modern science. It has certainly been one of its most important contributions to the current standard of living in all but the most undeveloped countries. Much of this is due to prevention through improved hygiene, and improved medical technology, particularly immunization and and treatment with modern antibiotics. However the reasons for changes in the prevalence of many infectious diseases are still poorly understood (tuberculosis in developed countries is an example)

and the economics of treatment and prevention of many diseases is still at a rudimentary stage of development. The formal basis for the study of epidemiology of infectious disease was only developed in the twentieth century. Pioneering work by Ross (1911), and Kermack and McKendrick (1927) established an empirically useful mathematical theory. Their work has been the foundation for almost all subsequent empirical and analytical research in epidemiology and cost-benefit analysis of control programs. Much of this research in reported in the authoritative references of Bailey (1975), and Anderson and May (1992).

Research on the costs and benefits of particular control programs has a long history, beginning with Daniel Bernoulli's analysis of smallpox prevention through variolation in the eighteenth century (See Bailey 1975 for a detailed discussion). There is now a large body of work on the cost-benefit analysis of infectious disease control in the public health literature. Macdonald (1965) is an early analysis of the control of helminth infections. Hethcote and Yorke (1984) is an analysis of the comparative effectiveness of different policies for control of gonorrhea. There is also a large literature on the control of tropical parasites, and childhood viral diseases such as measles, rubella and polio, as described in Anderson and May (1992). The full dynamic solution of the control problem is usually very difficult and a closed form solution does not exist except for very special cases. Therefore many dynamic analyses are numerical simulations of particular models. See Gupta and Rink (1973) for an example. This type of analysis, while useful for many public health issues, is lacking in robustness and does not yield general principles because the basic parameters of the disease processes vary, so the sensitivity and applicability of the results are open to question.

The economic analysis of the interaction of public health control strategies and individual incentives to engage in treatment and prevention activities is still in its infancy. The primary work in this area concerns the economics of immunization against vaccine preventable diseases, and HIV infection. Fine and Clarkson (1986) examine the difference between the social and individual private value of immunization levels. Brito, Sheshinski and Intriligator (1991) show that compulsory immunization is not optimal, and examine the optimal subsidies for immunization against vaccine preventable diseases in a population with heterogeneous tastes for engaging in prevention. Geoffard and Philipson (1992) study the incentives for eradication of vaccine preventable diseases. Philipson and Posner (1993, 1994) study the economics of prevention and treatment for HIV infection. Blower and MacLean (1994) and Le Point and Blower (1991) examine the effect of control

programs for HIV infection on the behavior of sexual active adults. These papers are important contributions to a poorly understood field. However, none of them provides a complete analysis of the dynamics and equilibrium of both the socially optimal policy for disease control and the effect of decentralized individually rational decision making by members the population. This is an important area of research because of the well publicized re-emergence of infectious disease as a serious public health threat, and the decreasing levels of funding for public health efforts. This paper attempts to provide the foundation for such an analysis.

The general analysis of the economics of infectious disease is quite difficult for at least four reasons. First, the generic lack of closed form solutions to the dynamic control problem, as mentioned above. This is related to the second reason: typically the value function for a social optimum will not be concave (See Lightwood 1994). Third, many individual decisions in the prevention and treatment of infectious disease are rather complex involving the dynamic nature of both costs and benefits. For example, immunization against childhood disease is an irreversible and individual specific durable investment rather than consumption of a flow service. A discussion of the difficulties involved in modeling economic decisions with regard to HIV infection is found in Philipson and Posner (1993). Four, infectious disease involves an obvious economic externality. An infected individual is often infectious, and not only incurs the cost of his own disease, but also imposes costs on other susceptibles in the population who are liable to be exposed to the contagion. This paper studies a simple model of medical treatment in one of the simplest models of infectious disease: the SIS model.

1.1.1. The SIS Infectious Disease Model Without Medical Treatment

The SIS infectious disease model is appropriate for a diseases for which both recovery and re-infection are likely to occur. Therefore the model is most often used for bacterial or parasitic infections for which no permanent immunity occurs after recovery. There are two states in the model: the susceptible state, S, and the infectious state, I. The letters S and I will be used to denote the state and the number of individuals in that state whenever there is no danger of confusion. A person in the infectious state S is healthy but susceptible to become infected with the disease upon exposure to the contagion. Upon infection the person enters the infected state I. The person remains in the infected state I until recovery, and is assumed to be infectious to susceptibles for the entire duration of the infected

state. There is a constant and finite probability of natural recovery in each period, and no superinfection is assumed to occur. The assumption of no superinfection limits the applicability of the model to microparasites such as bacteria, and certain macroparasitic diseases where the variation of the parasite population living in the individual is not important. Upon recovery the person re-enters the susceptible state. The initials SIS refer to the movement of a typical individual through the two states of the disease: Susceptible→Infectious→Susceptible.

The model will be in continuous time, and the individuals will be modeled as a continuum of representative agents. A large number approximation will be used for the probability of infection in this analysis, so the model will be deterministic and the infection rate will be used to model the continuous time probability rate of infection

The SIS model without medical treatment is usually presented in three equations:

1. $dS/dt = -(\beta I)S + \lambda I$
2. $dI/dt = (\beta I)S - \lambda I$
3. $S + I = N = 1$

where β is the *transmission coefficient* of the disease, lambda is the spontaneous recovery rate of an infected person, and S, I, β, $\lambda > 0$. For simplicity the total population N is normalized to unity.

Equations (2) and (3) are sufficient to characterize the disease process. The *force of infection* is equal to the transmission rate, β, multiplied by the level of infection in the population, I. The *incidence of infection*, βIS, results if the mass action principle of contagion holds. This means that infected persons spread a short lived infectious material into their local environment which then can be transmitted to susceptibles upon contact. The economic implication of the mass action principle is that modification of individual behavior is not a reliable means of prevention. This is particularly the case when the individual is infectious at or before the onset of the disease symptoms. Measles is a good example of a disease which exhibits this characteristic. People infectious with measles will spread the virus into the local environment when they breath. Susceptibles coming into contact with the same area will be exposed even after the infected person has been gone for up to several hours. The mass action principle does not hold in

general for sexually transmitted diseases, where direct person to person contact is required. See Anderson and Nokes (1991), or Cappasso (1993) for a more thorough discussion.

Setting $dI/dt = 0$ in (2), there are two possible sets of equilibrium levels of susceptibles and infectives, which are determined by the value of λ/β. If $\lambda/\beta > 1$ then there is one stable equilibrium,

$S^* = 1$, $I^* = 0$.

Disease can occur temporarily in this case if an infected is introduced to the population, however the epidemic will eventually die out. If $\lambda/\beta < 1$, there are two equilibria. There is one unstable equilibrium, often called the trivial equilibrium, where $I^* = 0$ and a stable equilibrium with a proper fraction of the population infected with the disease at all times,

$S^* = \lambda/\beta, I^* = 1 - \lambda/\beta$.

The unstable equilibrium will be broken by the introduction of a small number of infecteds and the disease process will approach the stable equilibrium above. All epidemics of the disease will approach this stable equilibrium and the disease is endemic.

Elementary introductions to the SIS model can be found in Hethcote (1976) and Allen (1994). Briscoe (1980) is a discussion of the use of the SIS model in early research on tropical parasites. A cost benefit analysis using the SIS model applied to the control of trachoma on a Native American Reservation can be found in Sanders (1971). Modifications of the SIS model can also be used as a basis for simple models of gonorrhea and syphilis, as shown in chapter 19 of Murray (1989).

1.1.2. The SIS Infectious Disease Model With Medical Treatment

This paper will consider the economics of a simple medical treatment. Our analysis begins with an extension of the work begun by Sanders (1971). Assume that during each period the infected individuals can purchase and consume medicine or other therapy that will increase their rate of recovery. The treatment will be assumed to be a flow good, and its effect is independent of the duration of treatment and it has no preventive properties upon recovery. Therefore only infecteds will purchase treatment, and their decision to purchase will be independent across time. The treatment will also be assumed to exist in discrete units, with each infected consuming exactly zero or one unit of treatment. Using equation (2) this results in the following model: $dI/dt = (\beta I)S - \lambda(I - M) - \delta M$, where $\delta > \lambda > 0$,

$I \geq M > 0$, and M is the number of infected individuals consuming treatment.

If $M = I$ in each period, i.e. every infected individual receives treatment then the model reduces to $dI/dt = (\beta I)S - \delta I$.

The equilibria of the model with full treatment is parallel to that of the model without treatment. If $\delta/\lambda < 1$ then there is an endemic full treatment equilibrium with one stable steady state with $S^* = \delta/\lambda$, $I^* = 1-\delta/\lambda$. Otherwise full treatment will eliminate the disease from the population.

Assume that the disease imposes a constant per period economic cost (say, in lost work days, reduced activity levels and physical suffering) of C_d, and a per period cost function for treatment of $C(M)$.The principal analysis below will be carried out assuming $C(M)$ exhibits increasing marginal cost, i.e. is convex, but a variety of alternative assumptions will also be considered.This completes the framework for the economic model.

It should be noted that in evaluating the benefit of treatment the infected will also need to evaluate the risks of re-infection in returning to the susceptible state. If the probability of re-infection is high, the benefit of treatment will be reduced. This is because, in the absence of any effective preventive measures, the infected will expect to avoid the costs of infection for a only short period since the individual expects to be re-infected quickly. However if the probability of infection is low, then the benefit of treatment will be higher because the infected will enjoy a long period in the susceptible state. Also note that at low levels of infection, full treatment can occur because the cost of treatment is low, but the benefits of treatment tend to be higher than otherwise because of the low probability of re-infection. At high levels of infection, it will often be the case that only a fraction of infected individuals will seek treatment because the cost of treatment will be high and quick re-infection is very likely.

Two assumptions on infecteds' expectations will be analyzed, including

1. static but continuously adjusted expectations, and

2. fully rational expectations in which the infected correctly predicts the entire future time path of the level and probability of infection in the population.

Sanders (1971) presents a similar *social* control model (using Calculus of Variations techniques) but restricts $C(M)$ to a simple linear form which results in a typical "bang-bang" solution for treatment. Subsequent work by Sethi (1974) (using Pontryagin's Maximum Principle) analyzes the same structure and employs

turnpike theory to consider the infinite horizon case. For linear costs, Sethi is able to establish the uniqueness of the optimal program. Our analysis extends these results in three significant directions:

1. The non-robust assumption of linear costs is removed and more general behavior emerges,

2. The dynamic adjustment paths are described and characterized,

3. The model examines individually rational behavior w.r.t. the socially optimal.

1.1.3. Limitations of the Model

The SIS model is a standard epidemiological model which has been found acceptable in empirical work. The practical application of the SIS model is limited compared to other mathematical models of the disease (e.g. the SIR model, which covers diseases for which infection confers permanent immunity, is applicable to such serious and important diseases as viral meningitis, hepatitis, and typhoid).Nonetheless, the SIS analysis serves as a foundation for other categories of disease.

The basic model is very simple and it ignores such factors as a incubation and latency periods, and the reduction or disappearance of infectiousness that often occurs before recovery. The assumption of mass action transmission has been found acceptable in empirical research for most diseases that do not require direct contact. A serious limitation of the model is the assumption of a constant and time independent recovery rate, with and without medical treatment. This assumption is most questionable for recovery with treatment, since in actuality repeated failure will often lead to a change in therapy, or an attempt at a rediagnosis. This is the same as assuming an exponential distribution of time to recovery, but it has been found to be acceptable as a rough approximation to time dependent recovery processes in epidemiology. See chapter 3 of Anderson and May (1992) for a fuller discussion.

The cost of the disease can be interpreted as the expected cost. Since the model is linear in terms of the cost of disease, this is an acceptable approximation for diseases which may have relatively rare but serious side effects for risk neutral agents. If the disease remains at a low level for a long period, there may be

delays in the application of treatment because the disease is unrecognized, or the appropriate protocol for treatment is poorly understood. In practice there also usually will be significant adjustment costs, which are ignored in this model.

2. The statement of the optimization problem.

Our problem then is to minimize the total discounted cost of disease - both direct and from treatment - over the indefinite future. We shall initially deal with the finite period version of this problem and then examine the limit of these programs as T becomes arbitrarily large.

2.1. The Maximum Principle[1]

Minimize the objective function:

$$\int_0^T \left(C_d I(t) + C(M(t))\right) e^{-\rho t} dt$$

where ρ is the rate of time preference, subject to the continuous time version of the infection equation:[2]

$$\frac{\partial I(t)}{\partial t} = \beta I(1 - I) - \lambda(I - M) - \delta M$$

2.1.1. Necessary Conditions for Optimization.

The spot value of the Hamiltonian expression for the intertemporal optimization problem can then be written

$$H = C(M) + C_d I + \varphi[\beta I(1 - I) - \lambda(I - M) - \delta M]$$

where the spot shadow cost of another infected individual, φ, changes according to:

$$\frac{\partial \varphi(t)}{\partial t} = \varphi\rho - \frac{\partial H}{\partial I}$$

[1]See e.g. Knowles (1981) especially Chapter 3 and 4.

[2]The time argument (t) is surpressed in the notation where no confusion would result.

or, expanding,

$$\frac{\partial \varphi(t)}{\partial t} = \varphi\rho - C_d + \varphi(-\beta + 2\beta I + \lambda)$$

and M is chosen to minimize H so

$$C'(M) + \varphi(\lambda - \delta) \left\{ \begin{array}{c} > \\ = 0 \\ < \end{array} \right\} \text{ only if } M = \left\{ \begin{array}{c} 0 \\ \in [0, I] \\ I \end{array} \right\}$$

With increasing marginal cost, M is an increasing function in φ[3]. The rationale for these diseconomies stem not only from production but from individual differences in the cost of obtaining treatment resulting from both personal impediments as well as difficulty of reaching increasingly more remote members of the effected population and, of course, crowding and capacity limitations in general. Finally, since $I(T)$ cannot reach zero in finite time, the transversality condition may be written $\varphi(T) = 0$.

2.1.2. The interpretation of φ.

Along the optimal path, the costate variable $\varphi(t)$ -the spot cost of another infected - equals the addition to the minimum present value of costs from t onward of another infected at time t. Minimizing the Hamiltonian w.r.t. M then trades off the instantaneous cost of treatment against the reduction in the *rate of change of infection* that results. For an interior solution this dictates that the marginal cost of treatment be equal to the marginal cost of infection, φ, multiplied by $(\delta - \lambda)$, the reduction in the *rate of change of infection* due to the additional treatment.

2.2. Characteristics of the Phase Space.

The $\frac{\partial I}{\partial t} = 0$ locus is described by

$$\beta I(1 - I) - \lambda(I - M) - \delta M = 0$$

or

$$-\beta I^2 + (\beta - \lambda)I = (\delta - \lambda)M$$

which describes a concave curve with intercepts of 0 and $\frac{\beta - \lambda}{\beta}$ in the IM phase space. Above this curve, I is decreasing and below it's rising. This locus is

[3]Sanders(1971) considers a special case where marginal costs are constant. We shall reconsider his model in a later section as a special case of our formulation here.

modified for implied values for M larger than I: When $I < \frac{\beta-\delta}{\beta}$ then M cannot attain a high enough value to cause $\frac{\partial I}{\partial t} = 0$ (since M is bounded by I). Thus, should $\frac{\beta-\delta}{\beta} > 0$, the $\frac{\partial I}{\partial t} = 0$ locus disappears for $I < \frac{\beta-\delta}{\beta}$ and further $\frac{\partial I}{\partial t} = 0$ for $I = \frac{\beta-\delta}{\beta}$ and $\varphi \geq \frac{C''(M)(\beta-\delta)}{\beta(\delta-\lambda)}$.

Now $\frac{\partial \varphi(t)}{\partial t}$ (or $\frac{\partial M(t)}{\partial t}$) is positive (or negative) as

$$C_d < (>)\varphi(\rho + 2\beta I + \lambda - \beta)$$

so the $\frac{\partial \varphi(t)}{\partial t} = 0$ locus is described by:

$$\varphi = \frac{C_d}{(\rho + 2\beta I + \lambda - \beta)}$$

When $I = 1$ this always yields a positive value for φ. As I is lowered the denominator falls and φ is higher rising asymptotically at $I = \frac{\beta-\rho-\lambda}{2\beta}$. For still smaller values of I, $\frac{\partial \varphi(t)}{\partial t}$ is always negative.

Letting $C''(M) = \alpha$, the steady states to $\frac{\partial I(t)}{\partial t} = 0$ and $\frac{\partial \varphi(t)}{\partial t} = 0$ must solve

$$((\beta - \lambda)I - \beta I^2)(\rho + \lambda - \beta + 2\beta I) - \frac{C_d(\delta - \lambda)^2}{\alpha} = 0$$

We shall characterize the Phase Space by its major structures - here the $\frac{\partial I(t)}{\partial t} = 0$ and $\frac{\partial \varphi(t)}{\partial t} = 0$ loci and the locus for the adjoint variable associated with full treatment, i.e. $M = I$. The stationary points are those associated with the intersection between $\frac{\partial I(t)}{\partial t} = 0$ and $\frac{\partial \varphi(t)}{\partial t} = 0$. Of course, $(\varphi, I) = 0$ is also a stationary solution in the sense that $\frac{\partial I(t)}{\partial t} = \frac{\partial M(t)}{\partial t} = 0$. It must be remembered that for particular parameter values the model may be degenerate and some, or even all, of these intersections may fail to occur. For our exposition here, we shall concentrate on the most general cases.

Figure 2.1 illustrates the nature of the phase space with the $\frac{\partial I(t)}{\partial t} = 0$ and $\frac{\partial \varphi(t)}{\partial t} = 0$ loci along with the $M = I$ boundary, on and above which the entire population is treated, indicated as a dashed line.[4]

[4]The diagrams are constructed using Maple V3 from the following parameter values: $\alpha = 1000$, $\beta = 0.12$, $\delta = 0.1$, $\lambda = 0.06$, $\rho = 0.05$, $C_d = 200$

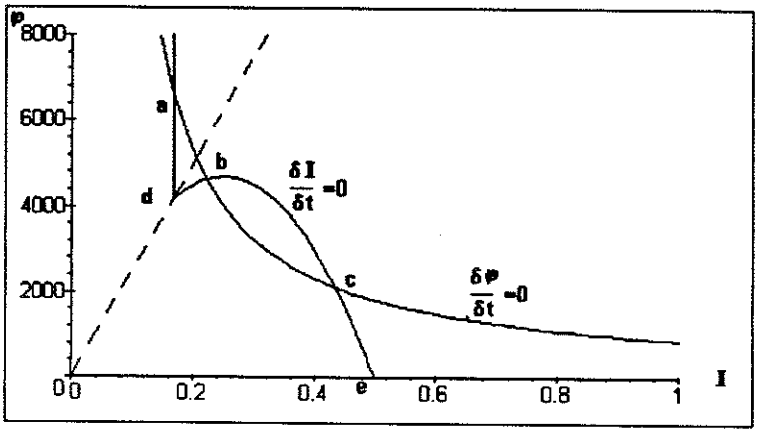

Figure 2.1:

2.2.1. An informal walk through the dynamics of the Phase Space

The $\frac{\partial\varphi(t)}{\partial t} = 0$ locus becomes asymptotic at $I = \frac{\beta-\lambda-\rho}{2\beta}$. For combinations of (I, φ) above this locus, φ is rising while below, it declines.

The $\frac{\partial I(t)}{\partial t} = 0$ locus has the form of an inverted parabola with I increasing below and decreasing above. There is, however, one modification which arises from the limitation that the number treated M cannot exceed the number infected, I. Thus for high values of the shadow price, φ, $\frac{\partial I(t)}{\partial t}$ becomes simply $\beta I(t)(1 - I(t)) - \delta I(t)$ so that when the total population of infected individuals is treated, I simply moves in the direction of $max\{0, \frac{\beta-\delta}{\beta}\}$. When $\frac{\beta-\delta}{\beta} > 0$, the $\frac{\partial I(t)}{\partial t} = 0$ locus becomes vertical at $max\{0, \frac{\beta-\delta}{\beta}\}$ and $\frac{\partial I(t)}{\partial t} = 0$ is undefined for $I \in (0, \frac{\beta-\delta}{\beta})$.

We may summarize the critical values for I in the phase space as follows:

1. the no-treatment equilibrium for I ((e) in figure 2.1): $\frac{\beta-\lambda}{\beta}$
2. the I asymptote for stationary φ: $I = \frac{\beta-\lambda-\rho}{2\beta}$
3. the full treatment saddle point ((a) in figure 2.1): $max\{0, \frac{\beta-\delta}{\beta}\}$
4. the low treatment saddle point ((c) in figure 2.1): the largest root to $((\beta - \lambda)I - \beta I^2)(\rho + \lambda - \beta + 2\beta I) - \frac{C_d(\delta-\lambda)^2}{\alpha} = 0$

5. the unstable steady state ((*b*) in figure 2.1): the smallest positive root to $((\beta-\lambda)I-\beta I^2)(\rho+\lambda-\beta+2\beta I)-\frac{C_d(\delta-\lambda)^2}{\alpha}=0$

2.2.2. The Stationary Loci

A necessary condition[5] that there be *interior* stationary solutions is $\beta-\lambda>0$ in which event there are as many as three stationary points in the phase space in addition to $(0,0)$:

1. A stationary point of type a exists if the $\frac{\partial\varphi(t)}{\partial t}=0$ locus intersects the vertical line at $I=\frac{\beta-\delta}{\beta}$ at a value of φ larger than $\lambda+\rho-\beta+2\beta\max\{0,\frac{\beta-\delta}{\beta}\}$(as in figure 2.1). Otherwise stated:

$$\frac{C_d}{\lambda+\rho-\beta+2\beta\max\{0,\frac{\beta-\delta}{\beta}\}}\geq\frac{\alpha\max\{0,\frac{\beta-\delta}{\beta}\}}{\delta-\lambda}$$

   The (possibly)[6] interior saddle point (a)describes full treatment at $M=I=\left(\frac{\beta-\delta}{\beta}\right)$ and $\varphi=\left(\frac{\beta-\delta}{\beta}\right)\left(\frac{\alpha}{\delta-\lambda}\right)$

2. Stationary points of type b and c exist if the unmodified $\frac{\partial I(t)}{\partial t}=0$ and $\frac{\partial\varphi(t)}{\partial t}=0$ loci intersect. This intersection will fail to occur if the cost of the disease, C_d, is sufficiently great. The solution (c) is at a relatively high infection rate characterized by a saddle point (the rightmost of the intersections between the $\frac{\partial I(t)}{\partial t}=0$ and $\frac{\partial\varphi(t)}{\partial t}=0$ loci), while the second (b) is described by an explosive and possibly cyclic point at a lower infection rate.(see the Appendix on the Equilibria in the Phase Space for the derivation of the roots to the approximating linear systems at the stationary points).

Since the optimization problem is not concave, the usual uniqueness and sufficiency characteristics of the transversality conditions fail and comparisons must be made along all paths satisfying the necessary (or first order) conditions.

The characterization of the optimal solution can now be described in terms of behavior w.r.t. these stationary points.

[5]By applying Descartes' rule of signs. In the event that $\beta-\lambda<0$, the disease will disappear on its own without intervention. The only economic question is what resources to use in hastening its inevitable self-eradication.

[6]The point a could occur on the vertical axis if $\frac{\beta-\delta}{\beta}\leq 0$.

2.3. The Finite Time Horizon

Since there is no constraint imposed on $I(T)$, then the optimal endpoint value for the adjoint variable along an optimal path, $\varphi^*(T) = 0$. As T becomes large, $\varphi(0)$ must adjust so as to lengthen the time it takes the trajectory to reach the horizontal axis. Figure 2.2 illustrates the backward paths from the I axis.In order

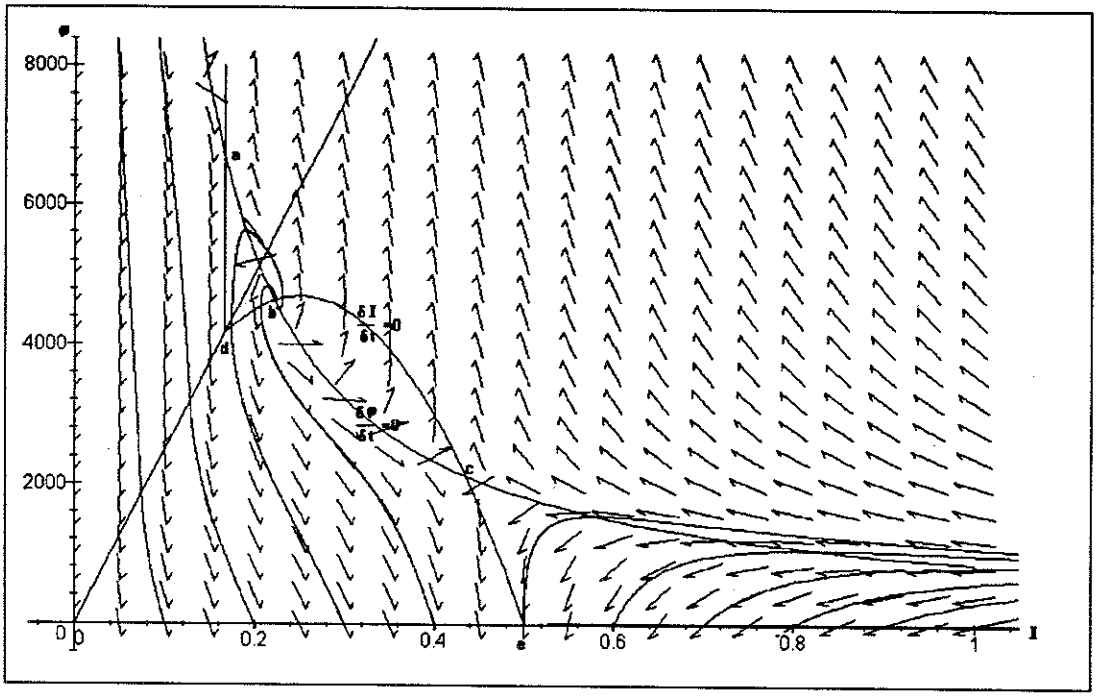

Figure 2.2:

to gain an understanding of the possible finite period optimal paths, consider all paths which terminate in finite time along the I axis. This may be accomplished by starting at points along that axis and running time backwards in the first order differential equations describing the motion in the phase space. Figure 2.2 illustrates this family of backward paths. In this illustration we note three distinct type of trajectories depending upon the terminal value for I. For low values, the backward paths miss the (a) saddle point to the left, for intermediate values they lead back to the unstable equilibrium (b) and for still higher values they trace

back toward the neighborhood of the saddle point at (c) to the right.

There are essentially four different stories (resulting from two possibilities in each of the following two cases) depending upon where the *backward* paths from the saddle points intersect the $\frac{\partial I(t)}{\partial t} = 0$ locus. (see figure 2.3 below for perhaps the most mathematically interesting).

1. If the left backward path from the rightmost saddle (c) does not intersect the $\frac{\partial I(t)}{\partial t} = 0$ locus to the right of $\max\{\frac{\beta-\delta}{\beta}, 0\}$ (point (d) in figure 2.1), the trajectory is monotone. The right backward path is always monotone and any initial $\varphi(0)$ below these paths results in $\varphi(t)$ reaching zero in finite time. For paths above, $\varphi(t)$ cannot converge to 0; it may approach a stable limit cycle or may become arbitrarily large. Thus, the backward paths from the I axis will lie under the two branches of this path. These optimal programs "ride along" below, the stable branch of the saddle point only to, eventually, depart downwards for the horizontal axis. As the time lengthens, these paths bow in toward the saddle point (in whose neighborhood they move with nearly negligible speed). These paths exhaust the possible solutions.

2. Alternatively, if the left backward path from the rightmost saddle (c) passes to the right of (d) , then the (reverse) stable branch bends backwards where it crosses the $\frac{\partial I(t)}{\partial t} = 0$ locus and spirals inward either toward (b) or to a stable limit cycle enclosing (b). Then for low initial values of I, paths which miss (a) on the low side may either proceed monotonically toward the I axis or even possibly spiral inward about (b) themselves. As time lengthens, these paths bow inward towards the turnpike at (a).

The branches to the two saddle points divide the space. Either one of the paths (both stable branches) misses the parabolic portion of the $\frac{\partial I}{\partial t} = 0$ locus or they both "connect" at (b). This latter case where both of the backward arms spiral in about (b) is depicted below in figure 2.3. This division will aid in defining a region for the $\varphi(0)$ where $\varphi(T)$ could reach 0.

2.4. Infinite Time Horizons

The choice of an *infinite* time horizon poses possible problems relating to existence of a solution. We shall instead suppose that there is a finite horizon, T, as above, then allow T to become arbitrarily large and examine the limiting path.

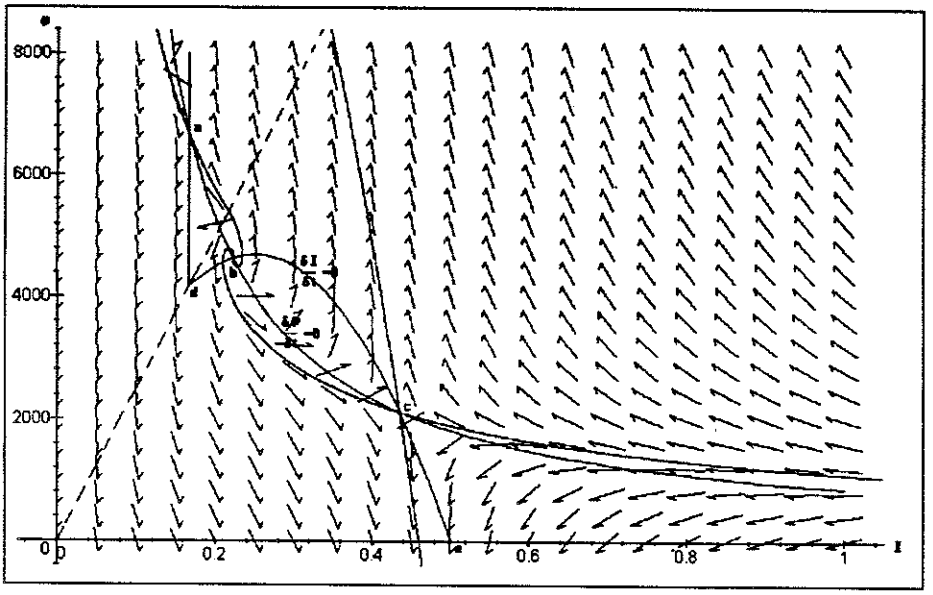

Figure 2.3:

Under the assumption that costs are quadratic, it is possible to show that any path which asymptotically approaches $I = \frac{\beta-\lambda}{\beta}$ *without* treatment is dominated and hence, cannot be optimal.[7] These are precisely the paths for which the shadow price becomes negative in finite time. This result follows directly from the assumption that marginal cost goes to zero as the level of treatment is small, so it would always be cost effective to treat minimally.

As $T \to \infty$ the optimal paths must take longer and longer to reach $\varphi(T) = 0$. The way in which this can be accomplished from a given starting $I(0)$ is to alter $\varphi(0)$ so that the paths lie closer to the stable branches from the saddle points causing the trajectories to "bow inward" toward those saddle points (where they move slowly) in the manner of the usual turnpike theorems in growth theory. Therefore the limiting program must either coincide with a stable branch to a saddle point converging to either the rightmost equilibrium (type (c))or else to the leftmost (type (a)).

[7]The argument is developed more fully in the Appendix on the Non-optimality of Inaction.

In the example above (figure 2.3), the limiting response as $T \rightarrow \infty$ is described by the stable branches to the saddle points at (a) and (c), i.e. the two spiral arms emenating from (b) outward to (a) and beyond and (c) and beyond. For any initial I, choosing $\varphi(0)$ to place oneself along this locus would lead toward one of the saddle points. However, where the paths are spirals, only those portions of the paths between the saddle points and the $\frac{\partial I(t)}{\partial t} = 0$ locus need be considered since it can never be optimal to choose an initial value for $\varphi(0)$ such that the trajectory returns to the same value of I at a later time (see figure 2.4).[8]

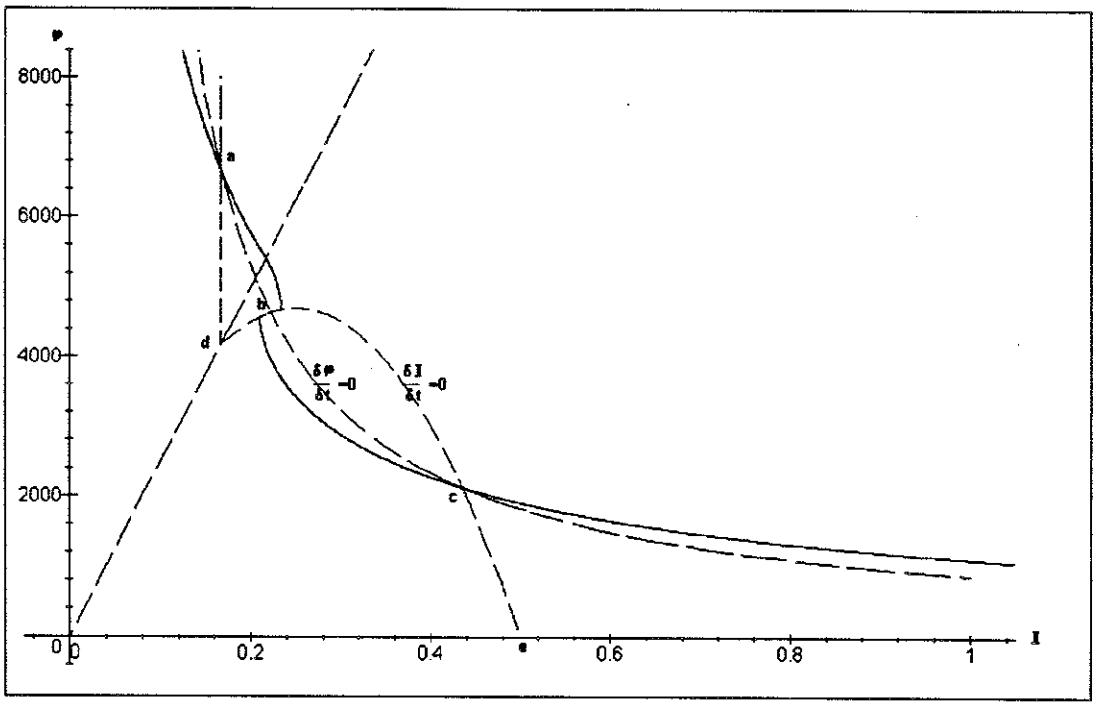

Figure 2.4:

The remaining overlap between these two paths cannot be resolved without direct computation - a consequence of the non-convexity in the problem itself. In that event that the rightward path does not extend fully to $I = 0$, there will be a discontinuity in the relationship between $I(0)$ and the optimal starting value for

[8]See the appendix on repetition of I values

$\varphi(0)$. There will be a single switching point (as we increase the starting value for I) where we "jump" from the leftward moving path to the rightward. [9]

Consequently, there is a negative monotone relationship between the level of infection and the optimal number of individuals treated - either of the two paths separately and even if there is a jump from the left path (i.e. the one through (a)) to the right one (through (c)). This somewhat surprising result is driven by the lowered value of treatment resulting from the increasing risk of reinfection as the number of infecteds in the population is increased.

2.5. A Brief Digression on Full Treatment and Eradication

2.5.1. Optimality and Full Treatment

Under what conditions could it *not* be optimal to *remain* in such a full treatment state? These circumstances are dealt with in the Appendix on Phase Space Behavior, and summarized as follows:

The left backward path from (c) must pass to the left of (d). Thus all finite period optimal program trajectories lie under the stable saddle point paths through (c) so their turnpike is at equilibrium (c). In this event, the limiting program must leave the full treatment state. This situation is guaranteed to arise when $\beta > \delta$ *and* $\rho + \lambda + \beta - 2\delta < 0$ (or $\beta < \delta$ *and* $\rho + \lambda - \beta < 0$). Eventually $\frac{\partial \varphi(t)}{\partial t}$ would become negative (and less than $-C_d$) in the neighborhood where I is near $max\{0, \frac{\beta-\delta}{\beta}\}$ and ultimately $\varphi(t) < 0$. As mentioned above, these programs cannot be efficient.

2.5.2. The Possibility of Eradication $\beta < \delta$:

The capability to eradicate of the disease entirely would require that $\frac{\partial I(t)}{\partial t}$ would be negative for small values of I. Now when $M = I, \frac{\partial I(t)}{\partial t} = [\beta(1 - I) - \delta] I$ so, for small I, $\frac{\partial I(t)}{\partial t} < 0$ if and only if $\beta - \delta \leq 0$.

[9] If the valuations are the same along the two paths for any two distinct values of I then, at intermediate I's, the valuations must also be the same. If this were not true, then it would be possible to start at a point on the lower valuation path, to then move along that path until the future valuations were equal and then jump to the other path and move in the reverse direction until the original value of I was reached again, thereby creating a superior (i.e. lower cost) cycle violating the necessary conditions. The argument is expanded in the Appendix on Valuation Paths.

This is exactly equivalent to the condition that the $I = M$ boundary (depicted in figure 1.1) lies everywhere above the $\frac{\partial I(t)}{\partial t} = 0$ locus. The slope of that locus, evaluated at $I = 0$, is equal to

$$\frac{(\beta - \lambda)\alpha}{(\delta - \lambda)^2}$$

while the slope of the boundary where $I = M$ is given by

$$\frac{\alpha}{\delta - \lambda}$$

We have presumed that both β and δ are larger than λ. If $\lambda > \beta$, the disease disappears without intervention.

2.6. Other Structures for Treatment Costs

2.6.1. Constant Marginal Costs

Where marginal costs are constant, say at MC, the above analysis takes on a slightly different flavor. The expression for $\frac{\partial \varphi(t)}{\partial t}$ along with the $\frac{\partial \varphi(t)}{\partial t} = 0$ locus remain unchanged but now $M = I$ (or 0) as $\varphi >$(or<) $\frac{MC}{\delta - \lambda}$. The phase space then takes on an extreme form of figure 2.1(see figure 2.5) with the same locus for $\frac{\partial \varphi(t)}{\partial t} = 0$ but with $\frac{\partial I(t)}{\partial t} = 0$ replaced by a horizontal line segment at $\frac{MC}{\delta - \lambda}$ between a vertical extension upwards at $\frac{\beta - \delta}{\beta}$ and one downwards at $\frac{\beta - \lambda}{\beta}$. These vertical segments denote the full treatment and no treatment equilibria respectively.When the shadow value of treatment $\varphi(t)$ exactly equals MC the level of treatment is indeterminant but could be set so as to equate $\frac{\partial I(t)}{\partial t}$ with 0.

The stability properties are as in the general case, with the intersection of the vertical segments and the $\frac{\partial \varphi(t)}{\partial t} = 0$ locus as saddle points and the intersection with the horizontal line as an unstable stationary state. As before, there are two sets of paths, one from each of the saddle points which constitute the possible solutions for optimal behavior. It is quite possible that an optimal program would initially treat and then cease treatment after the infection rate rose above some boundary. The turnpikes - or long run equilibria - would occur with either no or maximal treatment.

2.6.2. U-Shaped Average Costs

Our consideration here is principally with a program that incurs setup costs. The analysis is similar to that of the main presentation except that now treatment

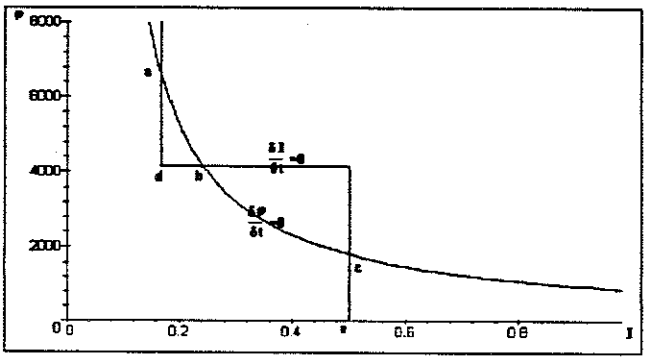

Figure 2.5:

ceases entirely when its shadow value, $\varphi(\delta - \lambda)$, falls below the minimum of average cost, MAC. The phase space picture in figure 2.1 is altered to reflect a region of zero treatment where $\varphi(\delta - \lambda) < MAC$. When φ attains this value, the level of treatment is at either zero or that associated with MAC. For values of φ above this level, the story is unaltered. The effect then is to split the phase space (see figure 2.6) horizontally at $\frac{MAC}{\delta - \lambda}$ and to replace the portion of the $\frac{\partial I(t)}{\partial t} = 0$ locus in the lower portion by a vertical line at $\frac{\beta - \lambda}{\beta}$. For $\varphi = \frac{MAC}{\delta - \lambda}$, $\frac{\partial I(t)}{\partial t}$ may be maintained at 0 by "chattering", though in the depiction in figure 2.6, $\frac{\partial \varphi(t)}{\partial t}$ would be negative and the system would quickly fall into the no-treatment state. [10]

2.6.3. Decreasing Marginal Costs

Where the marginal cost of treatment, $C'(M)$, is actually declining, then it is Hamiltonian minimizing to treat the entire infected group if $\frac{C(I)}{I} \leq \varphi(\delta - \lambda)$ and otherwise treat none. The horizontal portion of the $\frac{\partial I(t)}{\partial t} = 0$ locus - which separates the full treatment from the no treatment region - in figure 2.5 is here replaced by the montonically declining $\frac{C(I)}{(\delta - \lambda)I}$. There may be numerous special cases depending now on the additional complexity of the relative curvatures of

[10] It is possible to "pick up" two additional equilibria if the $\frac{\partial \varphi(t)}{\partial t} = 0$ locus intersects the new flat and vertical portions of the $\frac{\partial I(t)}{\partial t} = 0$ locus - one, an unstable equilibrium at the minimum of the average cost curve and the other a stable no-treatment equilibrium.

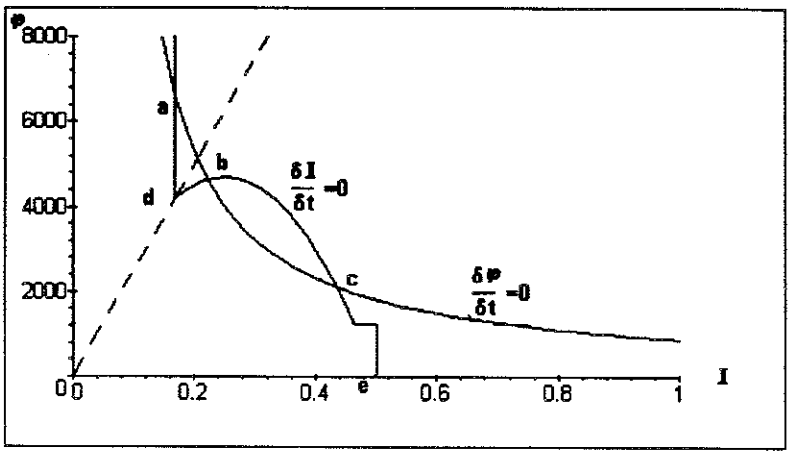

Figure 2.6:

the $\varphi = \frac{C(I)}{(\delta-\lambda)I}$ boundary and the $\frac{\partial\varphi(t)}{\partial t} = 0$ locus.

3. Individual Behavior

3.1. The behavior of firms:

Here, we shall be deliberately vague. The structural nature of the treatment "industry" is a potential source of concern and is a basis for study in itself. Decreasing marginal costs may lead to a natural monopoly while increasing marginal costs could result in a veritable continuum of firms.We will simply assume that each agent can purchase one unit of the treatment at the marginal cost of production as reflected by $C'(M)$.

3.2. Stationary Expectations

The individual's treatment decision depends upon forecasting the likelihood of future reinfection. Here, the rate of future infection in the population is taken as a stationary projection of the current rate. At each moment, the individual is presumed to change that projection to reflect the current rate of infection. We shall characterize an equilibrium in such a model as a forecast which, if believed,

would indeed instantly be realized from the cost minimizing individual behavior.

The individual's problem may then be reduced to treating if and only if

$$\int_0^\infty ((C_d + P)y_\tau(t))\, e^{-\rho t} dt \leq \int_0^\infty ((C_d)y_\mu(t))\, e^{-\rho t} dt$$

where $y_\tau(t)$ and $y_\mu(t)$ denote the probabilities of being infected at time t where the individual always chooses treatment or no treatment, respectively.

In particular, the probabilities of such individuals being infected at some future time are given by:

$$\frac{\partial y_\tau(t)}{\partial t} = \beta I(t)(1 - y_\tau(t)) - \delta y_\tau(t)$$

and

$$\frac{\partial y_\mu(t)}{\partial t} = \beta I(t)(1 - y_\mu(t)) - \lambda y_\mu(t)$$

where the change in the individual's probability of infection equals the likelihood of infection if healthy less that of cure if infected. Supposing that the societal rate of infection is perceived to be constant, these equations may be solved for $y_\tau(t)$ and $y_\mu(t)$ starting with an individual who is infected, i.e. $y(0) = 1$. Thus

$$y_\tau(t) = \frac{\beta I + \delta e^{-(\beta I + \delta)t}}{\beta I + \delta}$$

and

$$y_\mu(t) = \frac{\beta I + \lambda e^{-(\beta I + \lambda)t}}{\beta I + \lambda}$$

The integrals of these costs, treated and untreated, become, respectively,

$$\int_0^\infty ((C_d + P)y_\tau(t))\, e^{-\rho t} dt = \frac{(C_d + P)(\beta I + \rho)}{(\beta I + \rho + \delta)\rho}$$

and

$$\int_0^\infty ((C_d)y_\mu(t))\, e^{-\rho t} dt = \frac{C_d(\beta I + \rho)}{(\beta I + \rho + \lambda)\rho}$$

An equilibrium is described by a triple $\{P, I, M\}$ such that:

1. $P \leq \frac{C_d(\delta - \lambda)}{\beta I + \lambda + \rho}$ with strict inequality implying that $M = I$.
2. $\alpha M = P$

3. $\beta I(1-I) - \lambda(I-M) - \delta M = 0$

The first condition simply states that individuals will purchase treatment if the marginal benefit of treatment exceeds its price. The second states the equality of the marginal cost of treatment to its market price - the usual competitive condition and the third states the long run equilibrium condition that the overall level of infection in the society is constant at I.

Then, in an interior solution,

$$\alpha M = \frac{C_d(\delta - \lambda)}{\beta I + \lambda + \rho},$$

$$\beta I(1-I) - \lambda(I-M) - \delta M = 0.$$

Solving simultaneously, I must satisfy

$$C_d(\delta - \lambda)^2 - \alpha I(\beta - \lambda - \beta I)(\rho + \lambda + \beta I) = 0$$

There is a boundary solution at $I = M$, where $I = \max\{0, \frac{\beta - \delta}{\beta}\}$. These equilibria correspond to those of the more sophisticated expectations model below.

3.3. Individually Rational Behavior

Our model here corresponds to the notion of rational expectations. Individuals correctly predict the future *path* of the infection rate in the population and their optimizing behavior brings about this very path. Suppose first that the societal level of infection is perceived to follow some future path $I(t)$. The individual, if infected, bears a cost C_d and may choose treatment at some additional per period cost of P. In deciding whether or not to treat, the individual must look forward to the future probability of becoming reinfected. The higher that likelihood, the less valuable will be intervention. As before, the proposed treatment raises the rate of cure from λ to δ.

The representative individual's problem may then be reduced to minimizing

$$\int_0^\infty \left((C_d + \psi(t)P(t))y(t)\right) e^{-\rho t} dt$$

where $y(t)$ denotes the probability of being infected at time t, $P(t)$ is the cost of treatment at time t, and $\psi(t)$ is a control variable - (either 0 or 1) identifying

whether the individual chooses treatment if infected at time t. The probability of infection evolves as:

$$\frac{\partial y(t)}{\partial t} = \beta I(t)(1-y(t)) - \lambda(1-\psi(t))y(t) - \delta\psi(t)y(t)$$

where $I(t)$ is the rate of infection in the population as a whole. The problem is, again, one in the calculus of variations. The spot Hamiltonian expression governing this minimization is then

$$H = (C_d + \psi(t)P(t))y(t) + \varphi(t)[\beta I(t)(1-y(t)) - \lambda(1-\psi(t))y(t) - \delta\psi(t)y(t)]$$

where the future path of both $I(t)$ and $P(t)$ is presumed correctly known to the individual (i.e. perfect foresight). The control variable is chosen so as to minimize H thus $\psi(t) = 1$ if $P(t) + \varphi(t)[\lambda - \delta] > 0$ and $\psi(t) = 0$ if $P(t) + \varphi(t)[\lambda - \delta] < 0$. If we are to have an internal solution where some proper fraction of the infected population elects treatment, then $P(t) = \varphi(t)[\delta - \lambda]$. The shadow price $\varphi(t)$ must evolve according to $\frac{\partial \varphi(t)}{\partial t} = \varphi\rho - \frac{\partial H}{\partial y}$ or

$$\frac{\partial \varphi(t)}{\partial t} = -(C_d + \psi(t)P(t)) + \varphi(t)[\beta I(t) + \lambda + \rho - \lambda\psi(t) + \delta\psi(t)]$$

But since $P(t) = \varphi(t)[\delta - \lambda]$ this can be rewritten as

$$\frac{\partial \varphi(t)}{\partial t} = -C_d + \varphi(t)[\beta I(t) + \lambda + \rho]$$

Now in aggregate, the fraction of the population infected, $I(t)$ must be equal to the probability of any individual being in the infected state, or $y(t)$. Further, the number of individuals receiving treatment $M(t)$ must equal $\psi(t)y(t)$. [11] So, we have an equilibrium described by

1. $\alpha M(t) = \varphi(t)[\delta - \lambda]$
2. $\frac{\partial I(t)}{\partial t} = \beta I(t)(1 - I(t)) - \lambda I(t) - \frac{(\delta-\lambda)^2\varphi(t)}{\alpha}$.
3. $\frac{\partial \varphi(t)}{\partial t} = -C_d + \varphi(t)[\beta I(t) + \lambda + \rho]$

[11] Remember that with individuals indifferent to treatment or not, the fraction treated will be just that number necessary to equate the marginal cost of treatment with its shadow price.

The phase space characterizing the necessary conditions for an optimum trajectory has the same general appearance as in the socially optimal case. Indeed, the $\frac{\partial I(t)}{\partial t} = 0$ loci are identical but the $\frac{\partial \varphi(t)}{\partial t} = 0$ locus is now shifted downwards to

$$\varphi(t) = \frac{C_d}{\beta I(t) + \lambda + \rho}$$

- the denominator has increased from $\rho + 2\beta I + \lambda - \beta$ by $\beta(1 - I(t))$. The stationary solutions now are roots to

$$((\beta - \lambda)I - \beta I^2)(\lambda + \rho + \beta I) - \frac{C_d(\delta - \lambda)^2}{\alpha} = 0$$

and the rightmost saddle point is at a higher level of infection than in the socially optimal case.

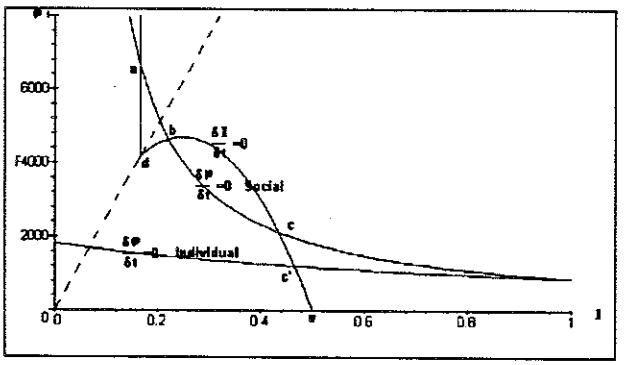

Figure 3.1:

3.4. A Non-marginal observation and the role of intervention

When, in the socially optimal model, the $\frac{\partial I(t)}{\partial t} = 0$ and $\frac{\partial \varphi(t)}{\partial t} = 0$ loci fail to intersect, the optimal path will ultimately involve full treatment. The individually rational model, with its lower $\frac{\partial \varphi(t)}{\partial t} = 0$ locus could still present a right saddle point solution with a higher asymptotic level of infection. *The difference is not merely the marginal shift from the higher locus but a global one* from the absence of the right saddle point in the social model. In this event, there may be a major

impact from the introduction of subsidies to modify individual behavior. In our example here, the only long run equilibrium in the individual case in figure 3.1 is at (c') and there is no individual solution corresponding to (a) at all! Since paths (individual or social) will spend most of there time near the turnpikes, for some initial conditions (e.g. for initial $I's$ in the neighborhood of (a)) the social optimum will be near (a) while the individually rational equilibrium will be located near (c'). Thus there can be a substantial duvergence between individually rational behavior and the social optimum.

The difficulty lies in the individual's failure to recognize the higher risk of infection which his or her own would impose on others. Thus, any single person undervalues the benefits of treatment. Standard solutions of the form of either subsidizing treatment or raising the cost of being infected (albeit a heartless approach) could serve to alter the individual incentives to correspond to the social costs.

3.5. Growing Population

If population, P, grows at some steady rate, say n, then the model may be recast in terms of $\frac{I}{P}$ and $\frac{M}{P}$. With two critical assumptions, the dynamics of the model remain virtually unaltered.

1. The rate at which susceptibles are infected depends upon the density of infecteds in the overall population, i.e. $\beta\frac{I}{P}$ replaces βI in the dynamics infection equation, and

2. The costs of treatment depends on both the level of treatment and the total population size in a linear homogeneous relationship, i.e. $C(M, P)$. The costs of treatment are assumed to vary from the distribution of individual characteristics in the population - e.g. variations in opportunity costs. Greater numbers are presumed to simply replicate this diversity.

In the dynamics, $\lambda + n$ replaces λ and the other variables, namely I and M are simply converted into per capita terms. This increase in the "effective" rate of recovery by n reflects the tendancy of $\frac{I}{P}$ to decline through growth in P.

4. Appendices

4.1. On the Non-optimality of Asymptotic Inaction

Consider any path for which $\varphi(t)$ eventually becomes negative. Several observations are in order. First, once negative, $\varphi(t)$ remains negative thereafter. Then $M(t)$ remains at zero and the level of infection grows to $\hat{y}= \frac{\beta-\lambda}{\beta}$. As shown here, under our assumptions regarding the costs of treatment, such a path cannot be optimal.

Suppose the path $\varphi(t)$ has reached zero by time τ and $y(t)$ is proceeding toward $\hat{y}$ and has reached $\hat{y} - \epsilon$. Redesignate this time as 0.

Then, uninterrupted, $\frac{\partial y}{\partial t} = \beta y(t)(1 - y(t) - \lambda y(t)$ and $y(0) = \hat{y} - \epsilon$. That is,

$$y(t) = \frac{-(\hat{y} - \epsilon)(\beta - \lambda)}{(\beta(\hat{y} - \epsilon) + \lambda - \beta)e^{-(\beta-\lambda)t} - \beta(\hat{y} - \epsilon)}$$

The costs of following this program are then

$$\int_0^\infty C_d \left[\frac{-(\hat{y} - \epsilon)(\beta - \lambda)}{(\beta(\hat{y} - \epsilon) + \lambda - \beta)e^{-(\beta-\lambda)t} - \beta(\hat{y} - \epsilon)} \right] e^{-\rho t} dt$$

Alternatively, we could remain at $\hat{y} - \epsilon$ by setting

$$M(t) = \frac{\beta(\hat{y} - \epsilon)(1 - (\hat{y} - \epsilon)) - \lambda(\hat{y} - \epsilon)}{(\delta - \lambda)}$$

and incur costs of

$$\int_0^\infty \left[C_d(\hat{y} - \epsilon) + 0.5\alpha \left(\frac{\beta(\hat{y} - \epsilon)(1 - (\hat{y} - \epsilon)) - \lambda(\hat{y} - \epsilon)}{(\delta - \lambda)} \right)^2 \right] e^{-\rho t} dt$$

The difference in costs is then

$$\Delta(\epsilon) = \int_0^\infty \left(\begin{array}{c} C_d((\hat{y} - \epsilon) \left[\frac{-(\beta-\lambda)}{(\beta(\hat{y}-\epsilon)+\lambda-\beta)e^{-(\beta-\lambda)t} - \beta(\hat{y}-\epsilon)} - 1 \right] \\ -0.5\alpha \left(\frac{\beta(\hat{y}-\epsilon)(1-(\hat{y}-\epsilon)) - \lambda(\hat{y}-\epsilon)}{(\delta-\lambda)} \right)^2 \end{array} \right) e^{-\rho t} dt$$

The first term in the above expression becomes

$$\frac{-(\beta-\lambda)-(\beta(\hat{\hat{y}}-\epsilon)+\lambda-\beta)e^{-(\beta-\lambda)t}+\beta(\hat{\hat{y}}-\epsilon)}{(\beta(\hat{\hat{y}}-\epsilon)+\lambda-\beta)e^{-(\beta-\lambda)t}-\beta(\hat{\hat{y}}-\epsilon)}=\frac{\beta\epsilon(e^{-(\beta-\lambda)t}-1)}{\beta\epsilon(1-e^{-(\beta-\lambda)t})-\beta+\lambda}$$

so the large integral expression becomes

$$\int_0^\infty\left(C_d\frac{-(\hat{\hat{y}}-\epsilon)\beta\epsilon(1-e^{-(\beta-\lambda)t})}{\beta\epsilon(1-e^{-(\beta-\lambda)t})-\beta+\lambda}-0.5\alpha\left(\frac{\beta(\hat{\hat{y}}-\epsilon)(1-(\hat{\hat{y}}-\epsilon))-\lambda(\hat{\hat{y}}-\epsilon)}{(\delta-\lambda)}\right)\right)e^{-\rho t}dt$$

Differentiating w.r.t. ϵ and setting ϵ equal to zero reveals that the *first term* becomes

$$\int_0^\infty C_d\frac{\begin{array}{c}(\beta\epsilon(1-e^{-(\beta-\lambda)t})-\beta+\lambda)\left(\left(-\beta\,\hat{\hat{y}}\,(1-e^{-(\beta-\lambda)t})+2\beta\epsilon(1-e^{-(\beta-\lambda)t})\right)\right.\\ \left.+\left((\hat{\hat{y}}-\epsilon)\beta\epsilon(1-e^{-(\beta-\lambda)t})\beta(1-e^{-(\beta-\lambda)t})\right)\right)\end{array}}{(\beta\epsilon(1-e^{-(\beta-\lambda)t})-\beta+\lambda)^2}$$

which becomes

$$\int_0^\infty C_d\frac{(1-e^{-(\beta-\lambda)t})\left((2\beta\epsilon-\beta+\lambda)(\beta\epsilon(1-e^{-(\beta-\lambda)t})-\beta+\lambda)+(\beta-\lambda-\beta\epsilon)\epsilon(1-e^{-(\beta-\lambda)t})\beta\right)}{(\beta\epsilon(1-e^{-(\beta-\lambda)t})-\beta+\lambda)^2}$$

or, for $\epsilon=0$,

$$\int_0^\infty C_d(1-e^{-(\beta-\lambda)t})e^{-\rho t}dt>0.$$

The derivative of the *second term* w.r.t. ϵ is:

$$-\int_0^\infty\alpha\left(\frac{\beta(\hat{\hat{y}}-\epsilon)(1-\hat{\hat{y}}+\epsilon)-\lambda(\hat{\hat{y}}-\epsilon))}{(\delta-\lambda)^2}\right)\left(-\beta(1-\hat{\hat{y}}+\epsilon)+\beta(\hat{\hat{y}}-\epsilon)+\lambda\right)e^{-\rho t}dt.$$

which reduces to:

$$-\int_0^\infty\alpha\left(\frac{\beta(\hat{\hat{y}}-\epsilon)(1-\hat{\hat{y}}+\epsilon)-\lambda(\hat{\hat{y}}-\epsilon))}{(\delta-\lambda)^2}\right)(-2\beta\epsilon+\beta-\lambda)\,e^{-\rho t}dt$$

When $\epsilon=0$ this becomes:

$$-\int_0^\infty\alpha\left(\frac{\beta\,\hat{\hat{y}}\,(1-\hat{\hat{y}})-\lambda\,\hat{\hat{y}})}{(\delta-\lambda)^2}\right)(\beta-\lambda)\,e^{-\rho t}dt$$

which, upon substituting $\hat{y} = \frac{\beta - \lambda}{\beta}$ can be shown equal to 0.

Combining these results we have shown that $\Delta'(0) > 0$ and since $\Delta(0) = 0$ it follows immediately that, for small ϵ, the cost of continuing to $\hat{y}$ will exceed that for remaining at $\hat{y} - \epsilon$. Since any path that converges to $\hat{y}$ will have to pass through this stage, it could not be optimal to discontinue treatment at any time prior to $\hat{y}$.

4.2. On the Equilibria in the Phase Space

The following treatment deals with the parabolic $\frac{\partial I(t)}{\partial t} = 0$ and the $\frac{\partial \varphi}{\partial t} = 0$ loci. The Hessian associated with the pair of differential equations is simply

$$\begin{bmatrix} -\beta + 2\beta I + \rho + \lambda & 2\beta\varphi \\ -\frac{(\delta - \lambda)^2}{\alpha} & -2\beta I + \beta - \lambda \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

At the rightmost stationary point where $\frac{\partial \varphi}{\partial I}$ along $\frac{\partial \varphi}{\partial t} = 0$ is negatively steeper than $\frac{\partial \varphi}{\partial I}$ along $\frac{\partial I}{\partial t} = 0$ or

$$\frac{-B}{A} > \frac{D}{-C}$$

respectively. (At the leftmost, this inequality is reversed.)

The eigenvalues must solve

$$(A - x)(D - x) - BC = x^2 - (A + D)x + AD - BC = 0$$

or

$$x = \frac{(A + D) \pm [(A + D)^2 - 4(AD - BC)]^{\frac{1}{2}}}{2}$$

If $BC > AD$ then the values of x are real and one root is positive and the other negative. If $BC < AD$ then the roots *may* be imaginary and exhibit cycles. Now at the rightmost equilibrium $\frac{\partial \varphi}{\partial I}$ along $\frac{\partial I}{\partial t} = 0$ and $\frac{\partial \varphi}{\partial I}$ along $\frac{\partial \varphi}{\partial t} = 0$ are negatively sloped so $0 > \frac{-B}{A} > \frac{D}{-C}$, and $-C$ and B are positive. Therefore A must also be positive and

$$BC > AD$$

roots are real and of opposite signs so the equilibrium is a saddle point.

At the leftmost equilibrium, the inequality relating the sloped of the stationary loci is reversed, i.e.

$$\frac{-B}{A} < \frac{D}{-C}$$

and $\frac{\partial \varphi}{\partial I}$ along $\frac{\partial \varphi}{\partial t} = 0$ is negatively sloped, so once again A and $-C$ are positive and now

$$BC < AD$$

and now x is either complex or has two positive real roots!

Finally, note that $A + D = \rho$ and therefore *if* the roots are complex, then the real part equals ρ and is positive! So behavior at the leftmost equilibrium is always explosive if $\rho > 0$, i.e. the rate of time preference is positive.

4.3. On Phase Space Behavior

Consider and differential equation $\frac{\partial f(t)}{\partial dt} = -C + \theta f(t)$. The general solution is then given by:

$$f(t) = \frac{C + \kappa\theta e^{\theta t}}{\theta}$$

where κ is determined by initial conditions.

Now the behavior of $\varphi(t)$ is bounded as follows:

$$-C_d+\varphi(t)(\rho+\lambda-\beta) \leq \frac{\partial \varphi(t)}{\partial dt} = -C_d+\varphi(t)(\rho+\lambda-\beta+2\beta I(t)) \leq -C_{d+}\varphi(t)(\rho+\lambda+\beta)$$

since I must lie in the interval $[0, 1]$.

But κ is determined by the initial conditions so (taking $\theta = (\rho + \lambda - \beta)$ and $C = C_d$) if $(\rho + \lambda - \beta) > 0$ then a choice of $\varphi(0) > \frac{C_d}{\rho+\lambda-\beta}$ will compel the corresponding κ to be positive and $\varphi(t)$ will go off to infinity. Eventually on such a path, $M(t)$ would become equal to $I(t)$ and remain so. Then $I(t) \to \max\{0, \frac{\beta-\delta}{\beta}\}$.

Alternatively, a choice of $\varphi(0) < \frac{C_d}{\rho+\lambda+\beta}$ would cause $\varphi(t)$ to approach $-\infty$.

For

$$\frac{C_d}{\rho+\lambda-\beta} > \varphi(0) > \frac{C_d}{\rho+\lambda+\beta}$$

or $(\rho + \lambda - \beta) < 0$ we need to examine the interaction between φ and I in more detail.

We are interested in whether, asymptotically, $\varphi(t)$ can remain high enough to cause $M(t) = I(t)$ indefinitely - the full treatment trajectory. Now $I(t) \rightarrow \max\{0, \frac{\beta-\delta}{\beta}\}$, so we have two cases to consider:

1. $\beta > \delta$: Supposing that we keep $M(t) = I(t)$ then $I(t) \rightarrow \frac{\beta-\delta}{\beta}$ and $\frac{\partial\varphi(t)}{\partial dt} \rightarrow -C_d + \varphi(t)(\rho + \lambda + \beta - 2\delta)$. If, initially $I(t) > \frac{\beta-\delta}{\beta}$, then $\frac{\partial\varphi(t)}{\partial dt} > -C_d + \varphi(t)(\rho + \lambda + \beta - 2\delta)$ so $\varphi(0) > \frac{C_d}{\rho+\lambda+\beta-2\delta}$ and $\rho + \lambda + \beta - 2\delta > 0$ imply the associated value for $\kappa > 0$ and $\varphi(t) \rightarrow \infty$. If $\rho + \lambda + \beta - 2\delta < 0$ then eventually $\varphi(t)$ must fall and continue to fall (when $I(t)$ nears $\frac{\beta-\delta}{\beta}$) and $M(t) < I(t)$.

2. $\delta > \beta$: The above argument is repeated except that $I(t) \rightarrow 0$ in the full treatment mode. Here $\frac{\partial\varphi(t)}{\partial dt} \rightarrow -C_d + \varphi(t)(\rho + \lambda - \beta)$. If $(\rho + \lambda - \beta) > 0$ then, for a choice of $\varphi(0) > \frac{C_d}{\rho+\lambda-\beta}, \varphi(t) \rightarrow \infty$. With $\rho + \lambda - \beta < 0$ then, again, eventually $\varphi(t)$ becomes low enough to cause $M(t) < I(t)$.

4.4. On repetition of I values

Suppose a path began at some value of I and φ and eventually returned to that same value of I at some later time τ. Denote the discounted cost during this interval as A and the discounted cost for the remainder of time to be B so that total cost is then

$$A + e^{-\rho\tau}B$$

If this program is better than simply starting with B then $A+e^{-\rho\tau}B < B$. But then, replacing the B in $A+e^{-\rho\tau}B$ by $A+e^{-\rho\tau}B$ would produce a still lower value, i.e. $A+e^{-\rho\tau}(A+e^{-\rho\tau}B)$ and so forth. This $A+Ae^{-\rho\tau}+Ae^{-2\rho\tau}... = \frac{A}{1-e^{-\rho\tau}}$ would be lower still. But this path is disontinuous in φ and violates the Pontryagin's necessary conditions.

4.5. On Valuation Paths

Consider the diagram in figure 4.1 respresenting the stable branches to the saddle points at (a) and (c).

Suppose that the valuations from I_1 and I_3 are equal but that the valuation from I_2 is lower on the bottom path. Let the time it takes to traverse region C be t_2, B be t_1, D be t_3 and E be t_4. Denote the valuation begining at I_3 and

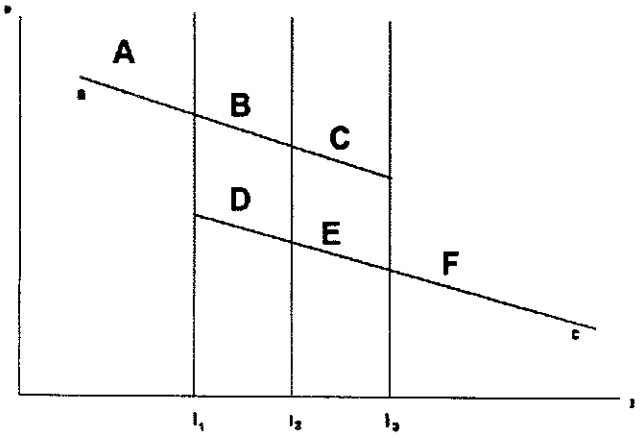

Figure 4.1:

proceeding along the upper path by $C+e^{-rt_2}B+e^{-r(t_1+t_2)}A$ and the valuation from I_2 along the lower path is strictly less than along the upper path, i.e. $E+Fe^{-rt_4} < B + Ae^{-rt_1}$. Then, multiplying both sides of this inequality by e^{-rt_2} and adding C we have

$$C + e^{-rt_2}E + Fe^{-r(t_2+t_4)} < C + e^{-rt_2}B + Ae^{-r(t_1+t_2)}$$

but the R.H.S. of this inequality is simply equal to F by assumption. Hence a contradiction.

5. Bibliography

Allen, LS (1994) Some Discrete Time SI, SIR and SIS Epidemic Models, Mathematical Biosciences 124. 83-105.

Anderson, RM. and RM. May (1992) Infectious Diseases of Humans: Dynamics and Control, Oxford, Oxford University Press.

Anderson, RM and DJ Nokes (1991) Mathematical Models of Transmission and Control, 225-252, chapter 14 in Oxford Textbook of Public Health, WW Holland, R Detels and G Knox, eds., Oxford, Oxford University Press.

Blower, SM. and AR MacLean (1994) Prophylactic Vaccines, Risk Behavior Change, and the Probability of Eradicating HIV in San Francisco, Science 265, 2 Sept., 1451-1454.

Briscoe, J (1980) On the Use of Simple Analytic Mathematical Models of Communicable Diseases, International Journal of Epidemiology 9, 265-270.

Brito, DL, E Sheshinski and MD Intriligator (1991) Externalities and Compulsory Vaccinations, Journal of Public Economics 45, 69-90.

Bailey, NTJ (1975) The Mathematical Theory of Infectious Diseases and Its Applications, London, Griffin.

Cappasso, V (1993) Mathematical Structures of Epidemic Systems, Berlin, Springer.

Fine, PEM and JA Clarkson (1986) Individual Versus Public Priorities in the Determination of Optimal Vaccination Policies, American Journal of Epidemiology 124, 1012-1020.

Geoffard, PY, and T Philipson (1992) Market Structure and Disease Eradication: Private vs. Public Vaccination, working paper, department of Economics, University of Chicago.

Gupta and Rink (1973) Optimum Control of Epidemics, Mathematical Biosciences 18, 383-396.

Hethcote (1976) Qualitative Analysis of Communicable Disease Models, Mathematical Biosciences 28, 335-356.

Kermack, WO, and AG McKendrick (1927) Contributions of the Mathematical Theory of Epidemics, Proceedings of the Royal Society A 115, 700-721.

Knowles, Greg (1981) An Introduction to Applied Optimal Control, New York, Academic Press.

Le Point, F, and SM Blower (1991) The Supply and Demand Dynamics of Sexual Behavior: Implications for Heterosexual HIV Epidemics, Journal of Acquired Immune Deficiency Syndromes 4,987-999.

Lightwood (1994) The Economics of Preventative Immunization: A Preliminary Analysis, unpublished manuscript,

Macdonald, G (1965) The Dynamics of Helminth Infections, with Special Reference to Schistosomes. Transactions of the Royal Society of Tropical Medicine and Hygiene 59, 489-506, reprinted in Modules in Applied Mathematics: Life Science Models, 1976, H. Marcus-Roberts and M. Thompson eds., Berlin, Springer.

Murray, JD (1989) Mathematical Biology, Berlin, Springer. Philipson, T and RA Posner (1993) Private Choices and Public Health: the AIDS Epidemic in and

Economic Perspective, Cambridge, MA, Harvard University Press.

Philipson, T and RA Posner (1994) Public Spending on AIDS Education: an Economic Analysis, Journal of Law and Economics 37, 17-38.

Ross, R (1911) The Prevention of Malaria, 2nd. ed., London, Murray.

Sanders,JL (1971) Quantitative Guidelines for Communicable Disease Control Programs, Biometrics 27, 833-893.

Sethi, S (1974) Quantitative Guidelines for Communicable Disease Control Program: A Complete Synthesis, Biometrics 30, 681-691.

Wickwire, K (1977) Mathematical Models for the Control of Pests and Infectuous Diseases: A Survey, Theoretical Population Biology 11, 182-238.

Wickwire, KH (1976) Optimal Conbtrol Policies for Reducing the Maximum Size of a Closed Epidemic - I. Deterministics Dynamics, Mathematical Biosciences 30, 129-137.

Wickwire, KH (1976) Optimal Conbtrol Policies for Reducing the Maximum Size of a Closed Epidemic - II. Stochastic Dynamics, Mathematical Biosciences 32, 1-14.