Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Health Economics 22 (2003) 575–598

JOURNAL OF HEALTH ECONOMICS

www.elsevier.com/locate/econbase

Resource allocation for control of infectious diseases in multiple independent populations: beyond cost-effectiveness analysis

Margaret L. Brandeau[a,*], Gregory S. Zaric[b], Anke Richter[c]

[a] *Department of Management Science and Engineering, Stanford University, Terman Building, Stanford, CA 94305, USA*
[b] *University of Western Ontario, London, ON, Canada*
[c] *Research Triangle Institute, Research Triangle Park, NC, USA*

Received 30 January 2002; accepted 13 January 2003

Abstract

Traditional cost-effectiveness analysis (CEA) assumes that program costs and benefits scale linearly with investment—an unrealistic assumption for epidemic control programs. This paper combines epidemic modeling with optimization techniques to determine the optimal allocation of a limited resource for epidemic control among multiple noninteracting populations. We show that the optimal resource allocation depends on many factors including the size of each population, the state of the epidemic in each population before resources are allocated (e.g. infection prevalence and incidence), the length of the time horizon, and prevention program characteristics. We establish conditions that characterize the optimal solution in certain cases.

JEL classification: C61; I19

Keywords: Optimization; Dynamic analysis; Resource allocation; Epidemics; Health

1. Introduction

Epidemics of infectious diseases afflict millions of people worldwide. For instance, more than 40 million people worldwide are infected with human immunodeficiency virus (HIV) (World Health Organization, 2001). Over 300 million cases of infection with sexually transmitted diseases other than HIV occur worldwide each year (World Health Organization,

* Corresponding author. Tel.: +1-650-725-1623; fax: +1-801-720-3610.
E-mail address: brandeau@stanford.edu (M.L. Brandeau).

0167-6296/03/$ – see front matter
doi:10.1016/S0167-6296(03)00043-2

1996). Some 8 million people become sick with tuberculosis each year, and 2 million people die (World Health Organization, 2000). If resources are unlimited, then the optimal way to allocate prevention funds is to spend enough money to reduce the disease transmission rate to its lowest possible value. When resources are constrained, however, the solution is no longer obvious. How should limited prevention funds be allocated in order to minimize the growth of an epidemic?

This paper combines epidemic modeling with optimization techniques to determine the optimal allocation of a limited resource for epidemic control. We consider a set of noninteracting populations, each described by a simple (susceptible/infected) epidemic model. Resources are to be spent to control transmission of the disease—in particular, to reduce the rate of sufficient contacts between susceptible and infected individuals. (The sufficient contact rate is defined as the average number of infecting contacts per infected person per unit time.) We assume that the relationship between the reduction in the sufficient contact rate in each population and the associated cost is described by a general cost function. The goal is to minimize the total number of people who become newly infected over the time horizon or to maximize the number of quality-adjusted life years gained, subject to a budget constraint and limits on attainable sufficient contact rates in each population.

Resource allocation has been studied by economists and operations researchers for many years (see for example (Hillier and Lieberman, 1993; Varian, 1978)), and much work has been done analyzing mathematical epidemic models (e.g. Anderson and May, 1991; Bailey, 1975). However, only limited work on optimal resource allocation for epidemic control has been done.

Cost-effectiveness analysis (CEA), a technique widely used to determine the worth of individual health care interventions, is often used for health care resource allocation. An incremental cost-effectiveness ratio (defined as the total incremental cost associated with an intervention divided by the total incremental benefits of the intervention) is calculated for each possible health care intervention. Benefits are usually measured in terms of quality-adjusted life years of survival so as to allow benefits for different medical conditions to be compared in a single unit of measure. Under the assumptions that interventions are mutually independent with no fixed start-up costs and with costs and health benefits that scale linearly, the optimal allocation of resources is achieved by selecting interventions in increasing order of their incremental cost-effectiveness ratios (Weinstein, 1995). However, epidemics tend to follow nonlinear growth curves, and incremental investment in an epidemic control program may not yield constant reductions in the chance of disease transmission, so the assumption of linear scaling of costs and health benefits may not be applicable to resource allocation problems involving infectious diseases. League tables allow decision makers to compare the cost-effectiveness ratios of several interventions. However, league tables frequently show only one level of program expansion, and therefore may be inappropriate for considering different levels of investment in competing epidemic control programs.

Stinnett and Paltiel (1996) presented several linear programming models that overcome these limitations (e.g. they allow for piecewise linear cost functions and linear scaling of cost functions following a fixed start-up cost). The goal is to maximize the sum of health benefits subject to a budget constraint. Such a framework has several drawbacks. The mixed integer-linear programming framework may not be applicable to epidemics with nonlinear, time-varying growth rates. Integer programming models are NP-hard (Garey and Johnson,

1979) and thus may be computationally intractable. In addition, such models offer little chance for developing insights regarding the optimal allocation of resources. Other linear and integer programming models for health care resource allocation problems have also been proposed (e.g. Chen and Bush, 1976; Gafni and Birch, 1993; van Zon and Kommer, 1999).

A number of researchers have applied control theory to simple epidemic models, assuming a single population (e.g. Blount et al., 1997; Greenhalgh, 1986; Greenhalgh, 1988; Muller, 1998; Sethi, 1974; Sethi, 1978; Sethi and Staats, 1978; Wickwire, 1977). Examples of controls typically considered include vaccination of susceptibles (Greenhalgh, 1986; Muller, 1998), treatment or removal of infectious persons (Greenhalgh, 1988; Sethi, 1978), and reduction in the sufficient contact rate between susceptibles and infectious persons (Blount et al., 1997; Sethi, 1974; Sethi and Staats, 1978; Wickwire, 1977). The goal is to determine the optimal control (for example, the optimal vaccination rate) over time. For analytical tractability, most analyses assume a single type of control (thus only one parameter is affected by the control). A typical objective might be to minimize the cost of control (e.g. immunization cost plus the fixed cost of establishing the immunization program) plus the cost associated with the number of individuals who become infected. With the exception of the fixed cost of establishing a control program, costs are usually assumed to be linear: the cost of control is a constant multiplied by the affected parameter, and the cost of disease is a constant multiplied by the number of individuals who become infected. Our work differs in two key ways: we consider multiple population subgroups, and we assume a general relationship between level of control and cost of control.

Another type of analysis considers allocation of resources among different population subgroups with the goal of disease eradication (reducing the disease equilibrium in each population to zero) (e.g. May and Anderson, 1984) or optimization of some function of the equilibrium state of the epidemic (e.g. Hethcote and Van Ark, 1987; May and Anderson, 1984). Our work assumes a finite time horizon, with the objective of minimizing the number of new infections that occur or maximizing the number of quality-adjusted life years gained.

Some researchers have used numerical analysis of sophisticated epidemic models to address resource allocation questions, but have not developed analytical results characterizing the optimal allocation of resources. One common approach applies simulation to compare a limited set of resource allocation alternatives (e.g. Bernstein et al., 1998; Hethcote, 1982; Hethcote and Yorke, 1984; Robinson et al., 1995). Another approach applies simulation and numerical analysis to consider all possible resource allocation alternatives (e.g. Longini et al., 1978; ReVelle et al., 1969; ReVelle and Male, 1970; Richter et al., 1999).

Many approaches to the resource allocation problem assume a linear relationship between funds expended and transmission reduction. For example, Kahn (1996) considered the impact of targeting HIV prevention resources to noninteracting populations. The analysis implicitly assumed linear cost functions, and considered only "all-or-nothing" solutions. Our analysis allows for a general relationship between funds expended and transmission reduction, and is not limited to all-or-nothing solutions.

Recent studies of resource allocation for epidemic control in multiple populations have incorporated more general cost functions for prevention programs. Kaplan (1998) and Kaplan and Pollack (1998) discussed a dynamic programming method for resource allocation. Richter et al. (1999) developed a model for allocating resources between two independent populations (injection drug users and nonusers) based on a framework developed by Richter

(1996), and Zaric and Brandeau (2001a) developed a model for allocating resources among several HIV prevention programs in a population of injection drug users and nonusers. Both models were solved numerically. Zaric and Brandeau (2001b) formulated a general model for allocating epidemic control resources among interacting populations and interventions over short time horizons. The problem is intractable even for small problem instances, so the authors developed approximations to the problem. They also developed some analytical results and approximation algorithms for a version of the problem in which epidemic control resources can be reallocated dynamically (Zaric and Brandeau, 2003).

In this work we shall use a simplified epidemic model and concentrate on developing theoretical results concerning the structure of optimal resource allocation policies. Previous work provided methodology for solving specific problem instances; the present work provides more general guidance on the resource allocation problem.

The remainder of this paper is organized as follows: we formulate the resource allocation problem in Section 2. In Section 3 we analyze the problem. We show that, in general, the problem is neither convex nor concave in the allocation variables. Moreover, the optimal resource allocation cannot always be simply stated as a function of a few problem parameters. Rather, the optimal resource allocation depends on the length of the time horizon, the size of each population, characteristics of the epidemic in each population before resources are expended, the cost functions for the prevention programs, and limits on attainable transmission rates. We present general results that characterize the optimal resource allocation in certain cases. We conclude with discussion in Section 4.

2. Problem statement

Consider an epidemic that exists in M distinct populations. These populations can also be considered as M distinct, noninteracting subgroups of a larger population. The epidemic within each population is described by a basic susceptible/infective (SI) epidemic model with two disease stages (susceptible and infected) and a constant population size (the rate of entry into the population equals the rate of exit from the population). The disease has no incubation period and, once infected, a person can transmit the disease continuously. There is no recovery from the disease. To specify the model, let

$$\lambda_i = \text{sufficient contact rate in population } i, \quad \lambda_i > 0$$

$$\delta_i = \text{replacement rate in population } i, \quad \delta_i > 0$$

$$S_i(t, \lambda_i) = \text{fraction of population } i \text{ that is susceptible at time } t, \text{ given sufficient contact rate } \lambda_i$$

$$I_i(t, \lambda_i) = \text{fraction of population } i \text{ that is infected at time } t, \text{ given sufficient contact rate } \lambda_i$$

for $i = 1, \dots, M$, $t \geq 0$. We assume that $S_i(t, \lambda_i) \in [0, 1)$ and $I_i(t, \lambda_i) \in (0, 1]$. The sufficient contact rate λ_i is defined as the average number of infecting contacts per infected person per unit time. That is, each infected individual will have on average λ_i contacts per unit time that may lead to infection. Assuming that the population is homogeneous and mixing uniformly, a proportion $S_i(t, \lambda_i)$ of those contacts will be with individuals who are not infected. Since there are $I_i(t, \lambda_i)$ infected individuals in the population at time t, the rate of infections per unit time equals $\lambda_i I_i(\cdot)S_i(\cdot)$ (Bailey, 1975). The replacement rate, δ_i, is the same for both infected and susceptible individuals and is assumed to be constant. Assuming the same exit rate for both infected and susceptible individuals is reasonable when either the probability of death due to the disease is small, or death due to infection is inevitable but occurs reasonably far into the future. The assumption of a constant replacement rate may be reasonable when the time horizon is relatively short, on the order of years rather than decades. The population size is constant, so

$$I_i(t, \lambda_i) + S_i(t, \lambda_i) = 1. \tag{1}$$

The differential equations that describe this model are (Bailey, 1975):

$$\begin{aligned} \frac{\mathrm{d}S_i}{\mathrm{d}t}(t, \lambda_i) &= -\lambda_i I_i(t, \lambda_i) S_i(t, \lambda_i) + \delta_i - \delta_i S_i(t, \lambda_i) \\ \frac{\mathrm{d}I_i}{\mathrm{d}t}(t, \lambda_i) &= \lambda_i I_i(t, \lambda_i) S_i(t, \lambda_i) - \delta_i I_i(t, \lambda_i) \end{aligned} . \tag{2}$$

We define $I_{0i} \equiv I_i(0, \lambda_i)$, and we assume that $I_{0i} \in (0, 1)$. The solution to Eq. (2) is (Bailey, 1975):

$$I_i(t, \lambda_i) = \begin{cases} \dfrac{(\lambda_i - \delta_i) I_{0i}\, \mathrm{e}^{(\lambda_i - \delta_i)t}}{I_{0i}\lambda_i(\mathrm{e}^{(\lambda_i - \delta_i)t} - 1) + (\lambda_i - \delta_i)} & \text{for } \lambda_i \neq \delta_i \\ \dfrac{I_{0i}}{I_{0i}\delta_i t + 1} & \text{for } \lambda_i = \delta_i. \end{cases} \tag{3}$$

The quantity $r_{0i} \equiv \lambda_i/\delta_i$ corresponds to the reproductive rate of infection in population i: it is the eventual number of new infections caused by one infected person. When $r_{0i} > 1$, the steady-state (equilibrium) infection prevalence in population i (as $t \to \infty$) is given by $1 - (1/r_{0i})$; when $r_{0i} \leq 1$, the prevalence goes to zero as $t \to \infty$.

Resources spent combating the epidemic are assumed to affect only the sufficient contact rate of the disease, and we assume that money spent on lowering the sufficient contact rate in one population does not affect the sufficient contact rate in another population. This latter assumption may not always be realistic: for example, television commercials aimed at one target population can also affect another population, and the correlation can be either positive or negative. However, due to the complexity of the correlations and the difficulty of establishing their values in a real situation, the assumption of independence is made.

We consider two possible objectives for the resource allocation problem: that of minimizing the number of new infections, and that of maximizing the number of quality-adjusted life years (QALYs) gained. Minimizing the number of new infections that occur is a natural objective for an epidemic control program; we also consider the objective of maximizing QALYs gained, as it is the recommended measure of benefit in applications of cost-effectiveness analysis to health and medicine (Gold et al., 1996). These quantities are measured over a given time horizon, from now, time zero, until T time units in the future ($T > 0$).

For each population i, the number of new infections that occur (calculated as the ratio of the number of new infections to the total population size) can be written as:

$$NI_i(\lambda_i) \equiv \int_0^T \lambda_i I_i(t,\lambda_i) S_i(t,\lambda_i)\,\mathrm{d}t = \int_0^T \frac{\mathrm{d}I_i(t,\lambda_i)}{\mathrm{d}t}\mathrm{d}t + \delta_i \int_0^T I_i(t,\lambda_i)\,\mathrm{d}t \tag{4}$$

The first term in (4) can be rewritten as:

$$\int_0^T \frac{\mathrm{d}I_i(t,\lambda_i)}{\mathrm{d}t}\mathrm{d}t = I_i(T,\lambda_i) - I_{0i} \tag{5}$$

where I_i (T,λ_i) is given by (3). Let

$$\begin{aligned} G_i(\lambda_i) &\equiv \delta_i \int_0^T I_i(t,\lambda_i)\,\mathrm{d}t \\ &= \begin{cases} \dfrac{\delta_i}{\lambda_i}\ln\left[\dfrac{(\lambda_i-\delta_i) + I_{0i}\lambda_i(\mathrm{e}^{(\lambda_i-\delta_i)T}-1)}{\lambda_i-\delta_i}\right] & \text{for } \lambda_i \neq \delta_i \\ \ln[\delta_i I_{0i} T + 1] & \text{for } \lambda_i = \delta_i \end{cases}. \end{aligned} \tag{6}$$

Substituting (5) and (6) into (4) yields the following expression for $NI_i(\lambda_i)$:

$$NI_i(\lambda_i) = I_i(T,\lambda_i) - I_{0i} + G_i(\lambda_i) \tag{7}$$

$$= \begin{cases} \dfrac{I_{0i}(\lambda_i-\delta_i)\,\mathrm{e}^{(\lambda_i-\delta_i)T}}{\lambda_i I_{0i}(\mathrm{e}^{(\lambda_i-\delta_i)T}-1) + (\lambda_i-\delta_i)} - I_{0i} & \\ \quad + \dfrac{\delta_i}{\lambda_i}\ln\left[\dfrac{(\lambda_i-\delta_i) + I_{0i}\lambda_i(\mathrm{e}^{(\lambda_i-\delta_i)T}-1)}{\lambda_i-\delta_i}\right] & \text{for } \lambda_i \neq \delta_i \\ \dfrac{I_{0i}}{\delta_i I_{0i} T + 1} - I_{0i} + \ln(\delta_i I_{0i} T + 1) & \text{for } \lambda_i = \delta_i \end{cases} \tag{8}$$

We define $\lambda \equiv (\lambda_1, \ldots, \lambda_M)$ and we let N_i denote the size of population i. The total number of newly infected people up to time T in the M populations is

$$\mathrm{INF}(\lambda) \equiv \sum_{i=1}^{M} N_i NI_i(\lambda_i). \tag{9}$$

The number of QALYs gained equals the number of QALYs experienced in the population when resources have been allocated to reduce the sufficient contact rates minus the number that would have been experienced with no resources allocated. The latter quantity is a constant, so the objective of maximizing the number of QALYs gained is equivalent to maximizing the number of QALYs experienced up to time T. We let $q_{i1} \in (0, 1)$ denote the quality adjustment for life years lived by susceptible individuals in population i and let $q_{i2} \in (0, q_{i1})$ denote the quality adjustment for life years lived by infected individuals in population i. We assume that quality of life is higher for susceptible individuals than for infected individuals; thus $q_{i1} > q_{i2}$. The number of QALYs experienced in population i

over the time horizon (calculated as the ratio of the number of QALYs experienced to the total population size) is

$$Q_i(\lambda_i) \equiv \int_0^T [q_{i1}S_i(t,\lambda_i) + q_{i2}I_i(t,\lambda_i)]\,dt \tag{10}$$

Substituting (1) into (10), we can rewrite $Q_i(\lambda_i)$ as

$$Q_i(\lambda_i) = q_{i1}T + \frac{q_{i2} - q_{i1}}{\delta_i} G_i(\lambda_i). \tag{11}$$

Eq. (11) is analogous to the expression for $NI_i(\lambda_i)$ in (7). The total number of QALYs experienced up to time T in the M populations is

$$Q(\lambda) \equiv \sum_{i=1}^{M} N_i Q_i(\lambda_i). \tag{12}$$

The goal of the resource allocation problem is to minimize (9) or maximize (12) subject to constraints on the attainable sufficient contact rates in each population and subject to a budget constraint. We assume that both upper and lower limits on the attainable sufficient contact rates exist. We let λ_{0i} denote the sufficient contact rate in population i at time 0, given no investment in the prevention programs. We assume that the lowest attainable sufficient contact rate in population i is a_i, where $0 < a_i < \lambda_{0i}$ (improvement is possible in every population). This lower limit could arise through resource constraints, considerations of epidemic virulence and modes of infection transmission, life spans of the infected, relative successfulness of interventions, and/or human behavior modification concerns and feasibilities.

We assume that it costs $c_i(\lambda_i)$ to attain sufficient contact rate λ_i in population i, and the total amount of available funds is B. Transient effects are not considered: in effect, the cost function $c_i(\lambda_i)$ is the net present cost of immediately achieving sufficient contact rate λ_i and maintaining that rate until time T. We assume that for each i, the function $c_i(\lambda_i)$ is continuous and differentiable for $a_i \le \lambda_i \le \lambda_{0i}$ and that $c_i(\lambda_{0i}) = 0$. The functions $c_i(\lambda_i)$ are assumed to be strictly decreasing in λ_i.

Examples of possible cost functions are shown in Fig. 1. The function in Fig. 1a corresponds to a prevention program with constant returns to scale: each incremental dollar spent on prevention yields the same incremental reduction in the sufficient contact rate. The function in Fig. 1b corresponds to a prevention program with decreasing returns to scale: for each incremental dollar spent on prevention, the corresponding reduction in the sufficient contact rate decreases. This may occur, for example, if individuals reached as a program expands are increasingly less likely to reduce their risky behavior. The function in Fig. 1c corresponds to a prevention program with increasing returns to scale: incremental investment in the prevention program leads to increasingly larger reductions in the sufficient contact rate. This may occur when a critical level of investment must be reached before significant levels of behavior change begin to accrue, or in situations where peer pressure plays a role in causing individuals to change their behavior. The cost curve of Fig. 1d incorporates both of these properties: the prevention program has increasing returns to scale for low levels of investment, but then for a large enough investment the program begins to

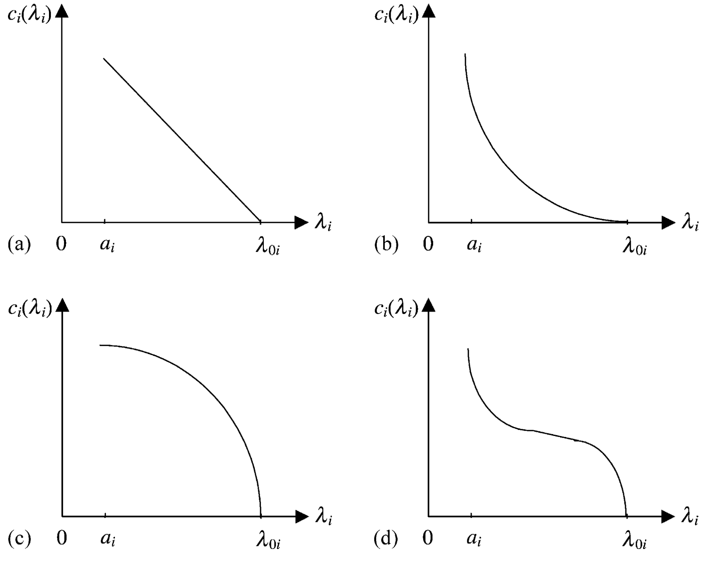

Fig. 1. Examples of possible cost functions $c_i(\lambda_i)$.

have decreasing returns to scale. Such a cost curve may be appropriate for many types of prevention programs. Other cost functions are also possible.

The budget constraint is written as:

$$\sum_{i=1}^{M} c_i(\lambda_i) \leq B. \tag{13}$$

When (13) holds as an equality, we say that the budget constraint is binding. We assume that $\sum_{i=1}^{M} c_i(a_i) > B$; the budget is not sufficient to reduce all transmission rates to their lowest achievable level.

With the above definitions, the resource allocation problem can be written as:

$$RA: \min_{\lambda_1,\ldots,\lambda_M} \mathrm{INF}(\lambda) \equiv \sum_{i=1}^{M} N_i NI_i(\lambda_i) \tag{9}$$

or

$$RA: \max_{\lambda_1,\ldots,\lambda_M} Q(\lambda) \equiv \sum_{i=1}^{M} N_i Q_i(\lambda_i) \tag{12}$$

$$\text{s.t.} \sum_{i=1}^{M} c_i(\lambda_i) \leq B \tag{13}$$

$$a_i \leq \lambda_i \leq \lambda_{0i} \quad i = 1, \ldots, M \tag{14}$$

3. Analysis of the resource allocation problem

We now analyze the problem *RA*. All proofs are in Appendix A. We first establish properties of the objective functions.

Lemma 1.

(i) $\mathrm{INF}(\underline{\lambda})$ *and* $Q(\underline{\lambda})$ *are continuous in* $\underline{\lambda}$ *and strictly positive*;
(ii) $\mathrm{INF}(\underline{\lambda})$ *is strictly increasing in* λ_i *for* $i = 1, \dots, M$;
(iii) $Q(\underline{\lambda})$ *is strictly decreasing in* λ_i *for* $i = 1, \dots, M$;
(iv) $\mathrm{INF}(\underline{\lambda})$ *and* $Q(\underline{\lambda})$ *are in general neither convex nor concave in* $\underline{\lambda}$.

Because of Lemma 1 (ii) and (iii) and our assumption that the budget is not sufficient to reduce all transmission rates to their lowest value, the budget constraint is binding in an optimal solution to *RA*. Thus, for the remainder of this paper, we will assume that the budget constraint is binding. We will refer to values of $\underline{\lambda}$ as the solution to the resource allocation problem, with the understanding that given λ_i, the associated level of expenditure on population i (the resource allocation) can be calculated from the function $c_i(\lambda_i)$.

In general $\mathrm{INF}(\underline{\lambda})$ and $Q(\underline{\lambda})$ may be convex, concave or neither. Examples for $\mathrm{INF}(\underline{\lambda})$ for the case of two populations (assuming linear cost functions, and assuming that the entire budget is spent) are shown in Fig. 2; examples for $Q(\underline{\lambda})$ are shown in Fig. 3. Below we will establish conditions under which convexity (concavity) of the two objective functions and the overall problem *RA* do hold.

To characterize the optimal solution to *RA*, we define Lagrange multipliers α, β_i, and γ_i, $i = 1, \dots, M$, and form the Lagrangian:

$$L(\underline{\lambda}, \alpha, \underline{\beta}, \underline{\gamma}) = \sum_{i=1}^{M} N_i J_i(\lambda_i) + \alpha\left[B - \sum_{i=1}^{M} c_i(\lambda_i)\right] + \sum_{i=1}^{M}[\beta_i(-a_i + \lambda_i)] + \sum_{i=1}^{M}[\gamma_i(\lambda_{0i} - \lambda_i)]$$

where $J_i(\lambda_i)$ denotes the objective function element $NI_i(\lambda_i)$ or $Q_i(\lambda_i)$. The first-order necessary conditions (FONCs) for optimality of *RA* are

$$\frac{\partial J_i(\lambda_i)}{\partial \lambda_i} - \alpha \frac{\partial c_i(\lambda_i)}{\partial \lambda_i} + \beta_i - \gamma_i = 0 \quad i = 1, \dots, M$$

$$\alpha\left[B - \sum_{i=1}^{M} c_i(\lambda_i)\right] = 0$$

$$\beta_i(-a_i + \lambda_i) = 0 \quad i = 1, \dots, M$$

$$\gamma_i(\lambda_{0i} - \lambda_i) = 0 \quad i = 1, \dots, M$$

$$\alpha, \beta, \gamma_i \geq 0 \quad i = 1, \dots, M$$

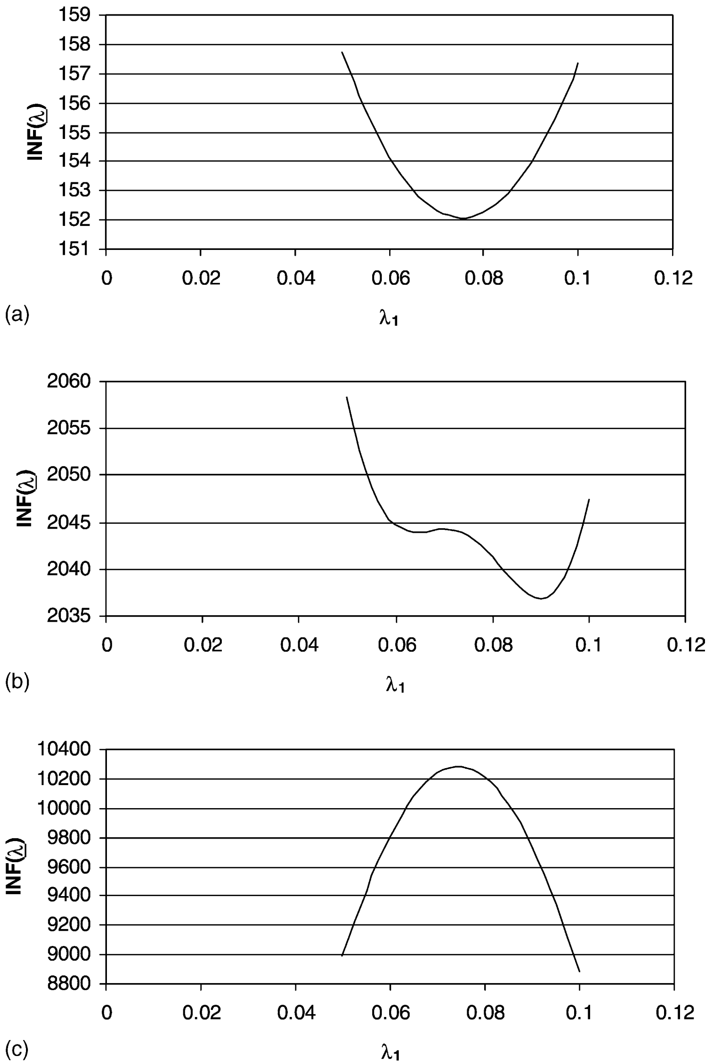

Fig. 2. Examples of INF(λ), assuming two populations and linear cost functions ($c_i(\lambda_i) = (0.1 - \lambda_i)/(N_i/0.010227)$), with $\lambda_2 = c_2^{-1}\,(B - c_1(\lambda_1))$ and B set so that $c_1(0.049) = B$, and with $N_1 = 1000$, $N_2 = 1010$, $\delta_1 = \delta_2 = 0.04$, $\lambda_{01} = \lambda_{02} = 0.1$, $I_{01} = I_{02} = 0.1$. For Fig. 2a, $T = 10$; For Fig. 2b, $T = 78$; for Fig. 2c, $T = 300$.

The optimal solution to *RA* satisfies these FONCs. However, finding the values of λ, α,$\{\beta_i\}$ and $\{\gamma_i\}$ that satisfy these conditions may be computationally difficult. One can use numerical methods to find local optima.

Although *RA* is based on one of the simplest possible types of epidemic models, the behavior of the objective functions as a function of λ is complex. In general, the optimal solution to *RA* cannot be characterized intuitively: the optimal solution depends on the length

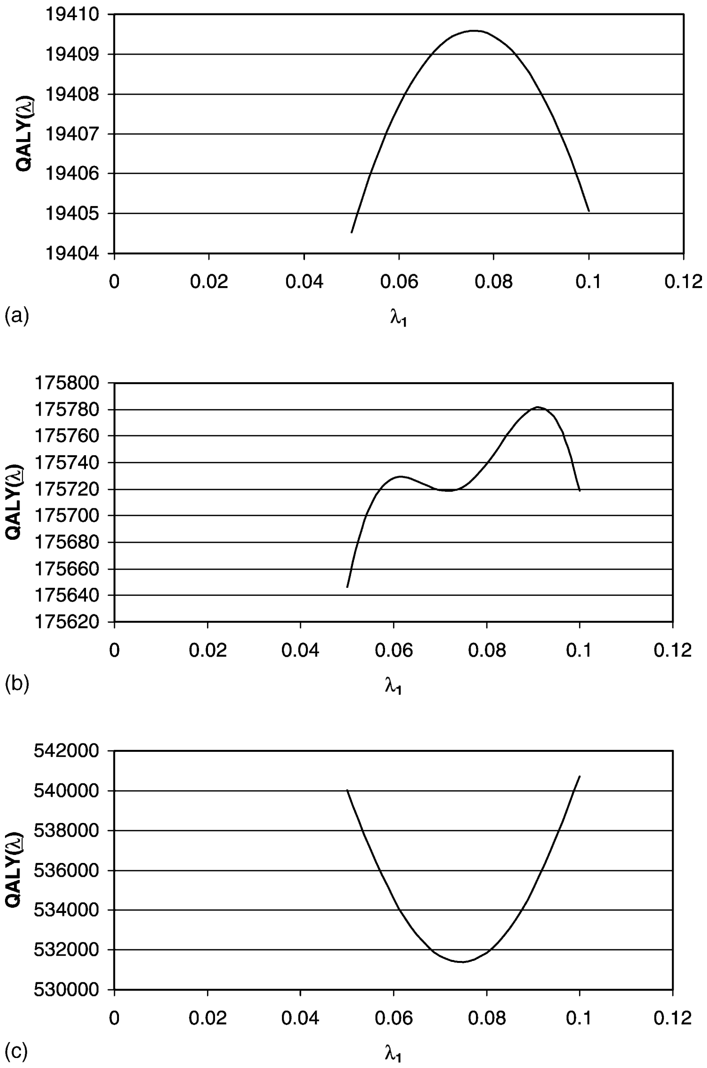

Fig. 3. Examples of Q(λ), assuming two populations and linear cost functions ($c_i(\lambda_i) = (0.1-\lambda_i)/(N_i/0.010227)$), with $\lambda_2 = c_2^{-1}$ $(B - c_1(\lambda_1))$ and B set so that $c_1(0.049) = B$, and with $N_1 = 1000$, $N_2 = 1010$, $\delta_1 = \delta_2 = 0.04$, $\lambda_{01} = \lambda_{02} = 0.1$, $I_{01} = I_{02} = 0.1$, $q_{11} = q_{21} = 1$, $q_{12} = q_{22} = 0.7$. For Fig. 3a, $T = 10$; for Fig. 3b, $T = 95$; for Fig. 3c, $T = 300$.

of the time horizon; the relative sizes of the populations; the initial infection prevalence, sufficient contact rate and replacement rate in each population; the cost functions; and limits on attainable transmission rates. For instance, in the example shown in Fig. 2a, the minimum value of INF(λ) occurs at $\lambda_1 = 0.075$. If $a_1 = 0.075$ (i.e. λ_1 cannot be reduced below 0.075), then it is optimal to invest as much of the budget as possible in Population 1, whereas if $a_1 < 0.075$, it is not optimal to invest as much as possible in Population 1. The example of Fig. 2a assumes $T = 10$. When the time horizon is longer (e.g. as in Fig. 2b and c), it is optimal to invest less in the program targeted to Population 1; the minimum value of INF(λ) occurs at a value larger than $\lambda_1 = 0.075$. For a long enough time horizon, it is optimal to invest as much as possible in the program targeted to Population 2. One can readily develop examples showing that it is not necessarily optimal to invest the most amount of money in the largest population, nor in the population with the largest initial sufficient contact rate, nor in the population with the highest initial infection prevalence. The examples of Figs. 2 and 3 assume specific linear cost functions. If the prevention program cost functions are different from those, the optimal allocations will also change.

We now present some results that characterize the optimal solution to *RA*—for both objective functions—when the time horizon T is sufficiently long. The following lemma provides the foundation for Theorems 1–3 below.

Lemma 2. *For T sufficiently large*:

(i) *$NI_i(\lambda_i)$ is concave in λ_i and $Q_i(\lambda_i)$ is convex in λ_i for $\lambda_i > \delta_i$;*
(ii) *$NI_i(\lambda_i)$ is convex in λ_i and $Q_i(\lambda_i)$ is concave in λ_i for $\lambda_i \in (\max(R_1, a_i), \min(R_2, \delta_i))$ where*

$$R_1 = \delta \frac{3(2 - I_0) - \sqrt{9I_0^2 - 4I_0 + 4}}{8(1 - I_0)}$$

$$R_2 = \delta \frac{3(2 - I_0) + \sqrt{9I_0^2 - 4I_0 + 4}}{8(1 - I_0)}.$$

Lemma 2 (i) says that over regions of λ_i for which the reproductive rate of infection in population i is greater than one, the total number of new infections in population i for T sufficiently large is a concave function of λ_i, and the total number of QALYs experienced is a convex function of λ_i. Lemma 2 (ii) establishes a region of λ_i for which, for T sufficiently large, $NI_i(\lambda_i)$ is convex in λ_i and $Q_i(\lambda_i)$ is concave in λ_i. We suspect, but were unable to prove analytically, that for T sufficiently large, $NI_i(\lambda_i)$ is convex in λ_i and $Q_i(\lambda_i)$ is concave in λ_i for all values of $\lambda_i < \delta_i$; however, we were unable to generate examples for $\lambda_i < \delta_i$ for which convexity of $NI_i(\lambda_i)$ and concavity of $Q_i(\lambda_i)$ did not occur for T sufficiently large.

The objective of maximizing QALYs (max $Q(\lambda)$) can be equivalently written as one of minimizing the negative of the QALYs function (min $-Q(\lambda)$). With this transformation, and the properties established in Lemma 2, it is clear that the convexity and concavity properties of the overall problem *RA* will be the same for the two objectives. Theorems 1–4 establish properties of *RA* and its optimal solution, for both objectives. The results are summarized in Table 1. We say that a point is an extreme point of the feasible region of solutions if it is not a linear combination of any two other points in the region.

Theorem 1. *For both objective functions, if the cost functions $c_i(\lambda_i)$ are convex in λ_i and if $a_i > \delta_i$ for $i = 1, \ldots, M$, then for T sufficiently large, the optimal solution to RA is an extreme point of the feasible region where the budget constraint is binding.*

Table 1
Summary of analytical results[a]

	Cost function characteristics	
Epidemic characteristics	Convex cost function (constant or decreasing returns to scale of λ_i with respect to c_i)	Concave cost function (constant or increasing returns to scale of λ_i with respect to c_i)
$a_i > \delta_i$ for all i Reproductive rate of infection greater than one in all populations Epidemic in each population has increasing returns to scale with respect to λ_i	Minimize a concave function over a convex region Theorem 1: optimal allocation is an extreme point of the feasible region; may involve sharing resources among populations	Minimize a concave function over a nonconvex region Theorem 2: optimal allocation is an all-or-nothing type of solution (invest as much as possible in certain populations)
$R_1 \le a_i < \lambda_{0i} < \min(R_2, \delta_i)$ for all i[b] Reproductive rate of infection less than one in all populations Epidemic in each population has decreasing returns to scale with respect to λ_i	Minimize a convex function over a convex region Theorem 3: optimal allocation is an extreme point of the feasible region; may involve sharing resources among populations	Minimize a convex function over a nonconvex region Optimal allocation cannot be easily characterized
$a_i > \delta_i$ for some i; $R_1 \le a_i < \lambda_{0i} < \min(R_2, \delta_i)$ for other i[b] Reproductive rate of infection greater than one in some populations, less than one in other populations Epidemic in each population has increasing returns to scale with respect to λ_i in some populations, and decreasing returns to scale with respect to λ_i in other populations	Theorem 4: allocate as much as possible to those populations with reproductive rate of infection greater than one; if funds remain, allocate to the other populations	

[a] All results assume a sufficiently long time horizon, and hold for both objective functions (minimizing the number of new infections or maximizing the number of quality-adjusted life years gained). An extreme point of the feasible region of solutions is a point that is not a linear combination of any two other points in the region.
[b] R_1 and R_2 are as defined in Lemma 2 (ii).

When Theorem 1 applies, one could solve *RA* by applying specialized algorithms for finding the minimum of a concave function over a convex region (e.g. Horst and Tuy, 1996). The result of Theorem 1 can be interpreted as follows: for a sufficiently long time horizon, when the number of new infections in any population i (or the number of QALYS experienced) exhibits increasing returns to scale with reductions in λ_i (which occurs when the reproductive rate of infection is greater than one), and the prevention programs have constant or decreasing returns to scale (i.e. convex cost functions), then the resource allocation problem is complex and depends on the specifics of the cost functions and the growth of the epidemic in each population. The optimal solution may be a greedy solution, but it may also involve sharing of resources among populations. A general intuitive statement of the optimal allocation of resources is difficult when Theorem 1 applies. Only in special cases will a greedy algorithm yield the optimal allocation of resources.

Theorem 2. *For both objective functions, if the cost functions $c_i(\lambda_i)$ are concave in λ_i and if $a_i > \delta_i$ for $i = 1, \dots, M$, then for T sufficiently large, the optimal solution to RA, λ^*, occurs at an extreme point of the feasible region where the budget constraint is binding, and satisfies $\lambda_i^* \in (a_i, \lambda_{0i})$ for at most one population i; for all other populations i, $\lambda_i^* = a_i$ or λ_{0i}.*

To find the optimal resource allocation when Theorem 2 applies, we can compare the values of $\mathrm{INF}(\lambda)$ (or $Q(\lambda)$) evaluated at the finite number of points of the feasible region of λ that satisfy the theorem statement and select as the solution the point that yields the smallest $\mathrm{INF}(\lambda)$ (or largest $Q(\lambda)$). An intuitive way of thinking about this result is as follows: for a sufficiently long time horizon, when the number of new infections in any population *i* (or the number of QALYS experienced) exhibits increasing returns to scale with reductions in λ_i (which occurs when the reproductive rate of infection is greater than one) and the prevention programs have constant or increasing returns to scale (i.e. concave cost functions), then an "all-or-nothing" type of solution is optimal: it is best to concentrate as much of the resource as possible in certain of the populations. It is straightforward to show, however, that the optimal solution to *RA* when Theorem 2 holds does not necessarily give as much of the resource as possible to the largest population; rather, the optimal solution depends on the growth of the epidemic in each population over the time horizon, the size of each population, and the cost functions.

Theorem 3. *For both objective functions, if the cost functions $c_i(\lambda_i)$ are convex in λ_i and if $R_1 \le a_i < \lambda_{0i} < \min(R_2, \delta_i)$ for $i = 1, \dots, M$ where R_1 and R_2 are as defined in Lemma 2* (ii), *then for T sufficiently large, RA is a convex optimization problem* (*that of minimizing a convex objective function over a convex feasible region*).

When Theorem 3 applies, the FONCs are sufficient for optimality. An intuitive way of thinking about this result is as follows: for a sufficiently long time horizon, when the number of new infections in any population *i* (or the number of QALYS experienced) exhibits decreasing returns to scale with reductions in λ_i (which occurs when $\mathrm{INF}(\lambda)$ is convex in λ_i and $Q(\lambda)$ is concave in λ_i), and the prevention programs have constant or decreasing returns to scale (i.e. convex cost functions), then an "all-or-nothing" type of solution may not be optimal; rather, an interior point solution, where resources are shared across populations, may be optimal.

Theorem 4. *For both objective functions, and for any type of continuous, strictly decreasing cost functions* $c_i(\lambda_i)$, $i = 1, \ldots, M$, *if* $a_i > \delta_i$ *for* $i = 1, \ldots, m_1$, *and if* $\lambda_{0i} < \delta_i$ *for* $i = m_1 + 1, \ldots, M$, *then for T sufficiently large, the optimal solution to RA satisfies the budget constraint as an equality and allocates no resources to populations* $i = m_1 + 1, \ldots, M$ *unless* λ_i *is reduced to* a_i *for all* $i = 1, \ldots, m_1$.

Theorem 4 says that, for sufficiently long time horizons, if some of the populations ($i = 1, \ldots, m_1$) have reproductive rate of infection greater than one, and some of the populations ($i = m_1 + 1, \ldots, M$) have reproductive rate of infection less than one, then it is optimal to invest in those populations with reproductive rate of infection less than one only if the maximum possible amount of money is first invested in the populations for which the reproductive rate of infection is greater than one. Intuitively, this occurs because the epidemic eventually dies out in populations for which the reproductive rate of infection is less than one, even when no resources are allocated to them, whereas in populations for which the reproductive rate of infection is greater than one, the epidemic continues indefinitely. Note that this is a long-term result; for shorter time horizons the optimal solution may involve sharing resources.

Theorems 1–3 provide insight into the optimal allocation of resources when the epidemic either has a nonzero equilibrium in all populations or when the epidemic eventually dies out in all populations. Note that when the reproductive rate of infection is less than one in all populations, and the cost functions have constant or increasing returns to scale, the optimal solution cannot be readily characterized, even for the case of a sufficiently long time horizon (see Table 1). Theorem 4 suggests that a reasonable heuristic for instances when some populations have a nonzero epidemic equilibrium and others have an epidemic that will eventually die out is to ignore the populations in which the epidemic will eventually die out and focus only on those populations for which the epidemic has a nonzero equilibrium.

Although the objective functions $\mathrm{INF}(\underline{\lambda})$ and $Q(\underline{\lambda})$ have similar convexity and concavity properties, the optimal allocations in general will be different for the two objectives. Theorem 5 establishes a condition under which an optimal solution to *RA* with the infections-averted objective function also maximizes the number of QALYs gained.

Theorem 5. *For any type of continuous, strictly decreasing cost functions* $c_i(\lambda_i)$, $i = 1, \ldots, M$, *if* $\delta_i = \delta$ *and* $q_{i1} - q_{i2} = q$ *for* $i = 1, \ldots, M$, *then for a sufficiently long time horizon T, the optimal solution to RA with the infections-averted objective is the same as the optimal solution with the QALYs gained objective*.

We note that when the conditions of Theorem 5 on $c_i(\lambda_i)$, δ_i, q_{i1}, and q_{i2} do not hold, the optimal solutions to *RA* for the two objectives are not necessarily the same, even for a sufficiently long time horizon.

4. Discussion

We have considered a problem of resource allocation for epidemic control among multiple independent populations where the objective is to minimize the total number of new infections that occur in the populations over a given time horizon or to maximize the total number of QALYs experienced. We have shown that the objective functions are not necessarily convex nor concave in the allocation variables and that the optimal resource allocation depends on many factors including the size of each population, the state of the epidemic in each population before resources are allocated, the length of the time horizon, and prevention program characteristics. We have also shown that both objective functions have similar convexity and concavity properties as a function of the sufficient contact rate.

We showed that the resource allocation problem has an all-or-nothing type of solution when the prevention programs have constant or increasing returns to scale (in reducing the sufficient contact rate), each population has nonzero epidemic equilibrium, and the time horizon is sufficiently long. The resource allocation problem is convex—leading to a readily found solution that may involve sharing the resource among the populations—when the prevention programs have constant or decreasing returns to scale (in reducing the sufficient contact rate), the epidemic equilibrium in each population is zero, and the time horizon is sufficiently long.

We have considered the allocation of resources to control an epidemic in multiple independent populations. Our model could also be used to allocate resources to control multiple epidemics in the same population, as long as the epidemics are independent. Not all epidemics of infectious disease are independent of one another: for example, some infectious diseases are related through similar risky behavior and exposure categories (e.g. hepatitis and HIV can both be spread through needle sharing (Lopez-Zetina et al., 2001)); and some infectious diseases are thought to be co-factors for other infectious diseases (e.g. some sexually transmitted diseases may be co-factors for HIV (Korenromp et al., 2001)). For epidemics that are not independent, a more sophisticated model would be needed to capture the interdependence among the epidemics.

Guidelines for interpreting cost-effectiveness ratios call for resources to be allocated in increasing order of cost-effectiveness ratios (Weinstein, 1995). However, the optimality of this solution relies on certain assumptions that do not hold for epidemics of infectious diseases, such as the assumption that costs and benefits scale linearly as a function of investment in a program. Such nonlinearities could be incorporated into a traditional cost-effectiveness analysis framework by performing CEA on small incremental expenditures. However, the results of such CEAs would represent piecewise linear approximations to the underlying nonlinear problem, and these approximations would probably need to be derived by considering the original nonlinear model. Furthermore, the common resource allocation algorithm of selecting interventions in increasing order of cost-effectiveness ratios would yield the best allocation only if all interventions had decreasing returns to scale (i.e. an investment of US$ 1000 after US$ 5000 has been invested is less cost effective than an investment of US$ 1000 when only US$ 2000 has been invested) over the entire feasible range of investments. If some interventions had increasing returns to scale, then the algorithm would not be valid. Results in this paper and elsewhere suggest that epidemic control programs may not exhibit strictly decreasing returns to scale. Nonetheless, CEA may be able to provide a reasonable approximation to the underlying nonlinear optimization problem, and its relative ease of use and interpretation may make CEA more appealing to decision makers than the formulation described in this paper. This motivates the development of similarly easy-to-use guidelines based on the formulation presented in this paper.

Our results (Theorems 1–4) show that the allocation of resources that minimizes the growth of an infectious disease epidemic is not necessarily a greedy solution. In some situations the optimal allocation may be "all-or-nothing". However, even in such situations, the optimal solution is not necessarily greedy in cost-effectiveness ratios. League tables allow policy makers to compare the relative cost-effectiveness of different interventions. The usefulness of league tables as a tool for determining the appropriate allocation of epidemic control resources is seriously questioned by the results in this paper. A recent report by the Institute of Medicine has advocated the use of a cost-effectiveness framework for allocating HIV prevention resources (Ruiz et al., 2001). Our work has shown that the nonlinear structure of epidemic models and cost functions leads to optimal solution properties that may differ from those found through naïve cost-effectiveness analysis or inspection of league tables. Given the magnitude of the impact that epidemics of infectious diseases have on the health and costs of health care for millions of people, it is crucial that policy makers use appropriate models for determining how to allocate resources to control infectious diseases.

Acknowledgements

Drs. Brandeau and Zaric were supported through grants from the National Institute on Drug Abuse (NIDA), National Institutes of Health (NIH) (Grants R01-DA09531, and R01-DA15612). Dr. Richter was supported by a VA Predoctoral Fellowship from the Department of Veterans Affairs. Dr. Zaric is the MBA '80 Faculty Fellow at the Ivey School of Business.

Appendix A

Proof of Lemma 1.

(i) Clearly $NI_i(\lambda_i)$ is continuous in λ_i for $\lambda_i \neq \delta_i$. It is straightforward to show by applying L'Hopital's rule to (7) that $NI_i(\lambda_i)$ is continuous at $\lambda_i = \delta_i$. Since INF(λ) is the sum of continuous functions, it is continuous. A similar argument holds for $Q(\lambda)$. It is well known for the SI model that $I(t, \lambda) \in (0, 1)$ for $I_0 \in (0, 1)$. Thus, for all i, $S_i(t, \lambda_i) = 1 - I_i(t, \lambda_i) > 0$ so $\lambda_i S_i(t, \lambda_i) I_i(t, \lambda_i) > 0$, and the integral (4) is also strictly positive. All terms in (10) are positive, so $Q_i(\lambda_i)$ and thus $Q(\lambda)$ are also strictly positive.

(ii) For ease of notation, we drop the subscript i. We first show that $\mathrm{d}I(t, \lambda)/\mathrm{d}\lambda > 0$, and then show that $\mathrm{d}NI(\lambda)/\mathrm{d}\lambda > 0$. From (3) we obtain

$$\frac{\mathrm{d}I(t,\lambda)}{\mathrm{d}\lambda} = \frac{I_0\,\mathrm{e}^{(\lambda-\delta)t}}{(\lambda I(\mathrm{e}^{(\lambda-\delta)t}-1)+(\lambda-\delta))^2}[(\lambda-\delta)^2 t + I_0\delta(\mathrm{e}^{(\lambda-\delta)t}-1) - I_0\lambda(\lambda-\delta)t], \quad \lambda \neq \delta \tag{A.1}$$

The first term in (A.1) (the ratio) is strictly positive, so we examine the second term. We denote the bracketed term by $A(t)$. $A(t)$ can be written as:

$$A(t) = A_1(t) + A_2(t) = [I_0\delta\mathrm{e}^{(\lambda-\delta)t}] + [t(\lambda-\delta)(\lambda-\delta-I_0\lambda) - I_0\delta] \tag{A.2}$$

We observe that $A_1(t) > 0$ for all $t > 0$, but $A_2(t)$ is not necessarily positive for all $t > 0$. The derivatives of $A(t)$ are:

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = [I_0\delta(\lambda-\delta)\,\mathrm{e}^{(\lambda-\delta)t}] + [(\lambda-\delta)(\lambda-\delta-I_0\lambda)] \tag{A.3}$$

$$\frac{\mathrm{d}^2A(t)}{\mathrm{d}t^2} = I_0\delta(\lambda-\delta)^2\,\mathrm{e}^{(\lambda-\delta)t} > 0 \quad \text{for all } t > 0 \tag{A.4}$$

Thus, $A(t)$ is convex in t. We observe from (A.2) that $A(0) = 0$, and observe that $A(t)$ is strictly increasing at $t = 0$ since

$$\left.\frac{\mathrm{d}A(t)}{\mathrm{d}t}\right|_{t=0} = [I_0\delta(\lambda-\delta)] + [(\lambda-\delta)(\lambda-\delta-I_0\lambda)] = (1-I_0)(\lambda-\delta)^2 > 0.$$

Thus, $\mathrm{d}I(t,\lambda)/\mathrm{d}\lambda > 0$. The derivative $\mathrm{d}NI(\lambda)/\mathrm{d}\lambda$ is

$$\begin{aligned}\frac{\mathrm{d}NI(\lambda)}{\mathrm{d}\lambda} &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(I(t,\lambda) - I_0 + \delta\int_0^T I(t,\lambda)\,\mathrm{d}t\right)\\ &= \frac{\mathrm{d}I(t,\lambda)}{\mathrm{d}\lambda} - \frac{\mathrm{d}I_0}{\mathrm{d}\lambda} + \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\delta\int_0^T I(t,\lambda)\,\mathrm{d}t\right)\end{aligned} \tag{A.5}$$

We showed above that the first term in (A.5) is positive. The second term is zero because I_0 is a constant. The third term is positive because it contains a positive integrand $(\mathrm{d}I(t,\lambda)/\mathrm{d}\lambda)$ over the range of integration. Thus, $\mathrm{d}NI(\lambda)/\mathrm{d}\lambda > 0$, and thus INF($\lambda$) is strictly increasing in λ_i for $i = 1, \ldots, M$.

(iii) For ease of notation, we drop the subscript i. From (11), we have

$$\frac{\mathrm{d}Q(\lambda)}{\mathrm{d}\lambda} = \frac{q_2-q_1}{\delta}\left[\frac{\mathrm{d}G(\lambda)}{\mathrm{d}\lambda}\right] \tag{A.6}$$

We proved in part (ii) that $\mathrm{d}G(\lambda)/\mathrm{d}\lambda > 0$. The term $(q_2 - q_1)/\delta$ is negative since $q_2 < q_1$, so the derivative (A.6) is strictly negative.

(iv) Examples for INF(λ) for the case of two populations (assuming identical linear cost functions for the two populations) are shown in Fig. 2; examples for $Q(\lambda)$ are shown in Fig. 3. □

Proof of Lemma 2. For ease of notation, we drop the subscript i. Lemma 3 below establishes that $G(\lambda)$ is concave in λ for $\lambda > \delta$ and convex in λ for $\lambda \in (\max(R_1, a), \min(R_2, \delta))$.

(i) $\lambda > \delta$: We first examine $N(\lambda)$, as expressed in (7). The term I_0 is a constant, and by Lemma 3, $G(\lambda)$ is concave. We examine the remaining term, $I(T,\lambda)$. Taking its derivative, we have

$$\frac{\mathrm{d}^2I(T,\lambda)}{\mathrm{d}\lambda^2} = \frac{k_0\,\mathrm{e}^{T(\lambda-\delta)} + k_1\,\mathrm{e}^{2T(\lambda-\delta)} + 2\delta I_0^3\,\mathrm{e}^{3T(\lambda-\delta)}}{(\delta-\lambda+I_0\lambda-I_0\lambda\mathrm{e}^{T(\lambda-\delta)})^3} \tag{A.7}$$

where k_0 and k_1 are terms that do not depend on T. As T becomes large, the dominating terms in (A.7) are those with exponent $\mathrm{e}^{3T(\lambda-\delta)}$; thus, as T becomes large, $\mathrm{d}^2I(T,\lambda)/\mathrm{d}\lambda^2 \to -2\delta/\lambda^3$, which is negative. Hence $N(\lambda)$ is convex for large T when

$\lambda > \delta$. We now examine $Q(\lambda)$, as expressed in (11). The term q_1T is strictly positive and linear in T. The term $[(q_2 - q_1)/\delta]G(\lambda)$ is convex for large T because it consists of a negative constant multiplied by a concave function. Thus $Q(\lambda)$ is convex for large T when $\lambda > \delta$.

(ii) $\lambda \in (\max(R_1, a), \min(R_2, \delta))$: We first examine $N(\lambda)$, as expressed in (7). The term I_0 is a constant. As $T \to \infty$, $I(T, \lambda) \to 0$ since $\lambda < \delta$. By Lemma 3, $G(\lambda)$ is convex for large T over the indicated region of λ. Thus, $N(\lambda)$ is convex for large T over the indicated region of λ. We now examine $Q(\lambda)$, as expressed in (11). The term q_1T is strictly positive and linear in T. The term $[(q_2 - q_1)/\delta]G(\lambda)$ is concave for large T because it consists of a negative constant multiplied by a convex function. Thus $Q(\lambda)$ is convex for large T over the indicated region of λ. □

Lemma 3. *For T sufficiently large*:

(i) *$G_i(\lambda_i)$ is concave in λ_i for $\lambda_i > \delta_i$;*
(ii) *$G_i(\lambda_i)$ is convex in λ_i for $\lambda_i \in (max(R_1, a_i), min(R_2, \delta_i))$ where R_1 and R_2 are as defined in* Lemma 2.

Proof of Lemma 3. For ease of notation, we drop the subscript i. The derivatives of $G(\lambda)$ are:

$$\frac{dG(\lambda)}{d\lambda} = \frac{\delta}{\lambda}\frac{1 + I_0(e^{T(\lambda-\delta)} - 1) + I_0\lambda T(e^{T(\lambda-\delta)} - 1)}{I_0\lambda(e^{T(\lambda-\delta)} - 1) + \lambda - \delta} - \frac{\delta}{\lambda^2}\ln\left(\frac{I_0\lambda(e^{T(\lambda-\delta)} - 1) + \lambda - \delta}{\lambda - \delta}\right) - \frac{\delta}{\lambda(\lambda - \delta)}$$

$$\frac{d^2G(\lambda)}{d\lambda^2} = \frac{2\delta}{\lambda^3}\ln\left(\frac{I_0\lambda(e^{T(\lambda-\delta)} - 1) + \lambda - \delta}{\lambda - \delta}\right) - \frac{\delta}{\lambda}\frac{[1 + I_0(e^{T(\lambda-\delta)} - 1) + I_0\lambda T\, e^{T(\lambda-\delta)}]^2}{[I_0\lambda(e^{T(\lambda-\delta)} - 1) + \lambda - \delta]^2} + \frac{\delta}{\lambda^2}\frac{-2 + 2I_0 - 2I_0\, e^{T(\lambda-\delta)} + I_0\lambda^2T^2\, e^{T(\lambda-\delta)}}{I_0\lambda(e^{T(\lambda-\delta)} - 1) + \lambda - \delta} + \frac{\delta(3\lambda - 2\delta)}{\lambda^2(\lambda - \delta)^2}$$

(i) $\lambda > \delta$: As $T \to \infty$, the first term in $d^2G(\lambda)/d\lambda^2$ is proportional to T, the second and third terms are each proportional to T^2, and the fourth is constant. Thus, we concentrate on the second and third terms. We find a common denominator and concentrate on the dominant terms, which are those containing the terms $T^2\, e^{2T}$:

$$\frac{\delta}{\lambda^2}\frac{-2 + 2I_0 - 2I_0\, e^{T(\lambda-\delta)} + I_0\lambda^2T^2\, e^{T(\lambda-\delta)}}{I_0\lambda(e^{T(\lambda-\delta)} - 1) + \lambda - \delta} - \frac{\delta}{\lambda}\frac{[1 + I_0(e^{T(\lambda-\delta)} - 1) + I_0\lambda T\, e^{T(\lambda-\delta)}]^2}{[I_0\lambda(e^{T(\lambda-\delta)} - 1) + \lambda - \delta]^2}$$

$$\approx \frac{\delta}{\lambda^2}\frac{I_0\lambda^2T^2\,\mathrm{e}^{T(\lambda-\delta)}[I_0\lambda(\mathrm{e}^{T(\lambda-\delta)}-1)+\lambda-\delta]}{[I_0\lambda(\mathrm{e}^{T(\lambda-\delta)}-1)+\lambda-\delta]^2}$$
$$-\frac{\delta\lambda}{\lambda^2}\frac{[I_0\lambda T\,\mathrm{e}^{T(\lambda-\delta)}]^2}{[I_0\lambda(\mathrm{e}^{T(\lambda-\delta)}-1)+\lambda-\delta]^2}$$
$$\approx \frac{\delta}{\lambda^2}\frac{I_0^2\lambda^3T^2\,\mathrm{e}^{2T(\lambda-\delta)}}{[I_0\lambda(\mathrm{e}^{T(\lambda-\delta)}-1)+\lambda-\delta]^2}-\frac{\delta}{\lambda^2}\frac{I_0^2\lambda^3T^2\,\mathrm{e}^{2T(\lambda-\delta)}}{[I_0\lambda(\mathrm{e}^{T(\lambda-\delta)}-1)+\lambda-\delta]^2}=0$$

Since the coefficient of terms containing $T^2\,\mathrm{e}^{2T}$ is zero, we now examine the term containing $T\,\mathrm{e}^{2T}$, which is found only in the second term of the expression for $\mathrm{d}^2G(\lambda)/\mathrm{d}\lambda^2$. This term is:

$$-\frac{\delta}{\lambda}\frac{2I_0^2\lambda T\,\mathrm{e}^{2T(\lambda-\delta)}}{[I_0\lambda(\mathrm{e}^{T(\lambda-\delta)}-1)+\lambda-\delta]^2}$$

which is clearly negative. Thus, $G(\lambda)$ is concave for large T.

(ii) $\lambda \in (\max(R_1, a), \min(R_2, \delta))$: As $T \to \infty$, all terms of the form $\mathrm{e}^{T(\lambda-\delta)} \to 0$ since $\lambda < \delta$. Thus, for large T,

$$\frac{\mathrm{d}^2G(\lambda)}{\mathrm{d}\lambda^2} \approx \frac{2\delta}{\lambda^3}\ln\left(\frac{\lambda-\delta-I_0\lambda}{\lambda-\delta}\right)-\frac{\delta}{\lambda}\frac{[1-I_0]^2}{[\lambda-\delta-I_0\lambda]^2}+\frac{\delta}{\lambda^2}\frac{-2+2I_0}{-I_0\lambda+\lambda-\delta}+\frac{\delta(3\lambda-2\delta)}{\lambda^2(\lambda-\delta)^2} \tag{A.8}$$

The coefficient of the first term of (A.8) is positive. Moreover,

$$\ln\left(\frac{\delta-\lambda+I_0\lambda}{\delta-\lambda}\right)>0 \quad \text{since} \quad \frac{\delta-\lambda+I_0\lambda}{\delta-\lambda}>\frac{\delta-\lambda}{\delta-\lambda}=1$$

which follows from the fact that $0 < I_0 < 1$. Thus, the first term in (A.8) is positive. We now examine the remaining terms in (A.8). We have:

$$-\frac{\delta}{\lambda}\frac{[1-I_0]^2}{[-I_0\lambda+\lambda-\delta]^2}+\frac{\delta}{\lambda^2}\frac{-2+2I_0}{-I_0\lambda+\lambda-\delta}+\frac{\delta(3\lambda-2\delta)}{\lambda^2(\lambda-\delta)^2}$$
$$=-\frac{\delta^2I_0[4(1-I_0)\lambda^2+3\lambda\delta(I_0-2)+2\delta^2]}{\lambda^2(\lambda-\delta)^2(\delta-\lambda+I_0\lambda)^2} \tag{A.9}$$

The denominator of (A.9) is clearly positive. The term in square brackets in the numerator of (A.9) is a quadratic equation in λ. The quadratic equation has two distinct real roots, since the discriminant is positive. The discriminant is given by:

$$[3\delta(I_0-2)]^2-4[4(1-I_0)][2\delta^2]=\delta^2[9I_0^2-4I_0+4]>0 \quad \text{since} \quad 0<I_0<1.$$

The two roots of the quadratic equation, R_1 and R_2, are specified in the statement of Lemma 2 (note that $R_1 < R_2$ since $2-I_0 > 0$ and $1-I_0 > 0$). The coefficient of λ^2 in (A.9) is positive. Thus, the term in square brackets in the numerator of (A.9) is negative for $R_1 < \lambda < R_2$ and positive otherwise. Since the entire expression is multiplied by a negative number, these terms are positive for $R_1 < \lambda < R_2$. Thus (A.8) is positive for $R_1 < \lambda < R_2$, and thus $G(\lambda)$ is convex. □

Proof of Theorem 1. The budget constraint is binding in an optimal solution to *RA*. Since $a_i > \delta_i$ by assumption, we have $\lambda_i > \delta_i$ everywhere in the feasible region of $\underline{\lambda}$. By Lemma 2, $NI_i(\lambda_i)$ and $-Q_i(\lambda_i)$ are concave in λ_i over $\lambda_i > \delta_i$ for T sufficiently large. Since positive multiples and sums of concave functions are concave, INF($\underline{\lambda}$) and $-Q(\underline{\lambda})$ are concave in λ_i. The two objective functions are separable in λ_i, so INF($\underline{\lambda}$) and $-Q(\underline{\lambda})$ are concave in $\underline{\lambda}$. Convex cost functions and the linear constraints on $\underline{\lambda}$ lead to a convex feasible region. The minimum of a concave function over a convex feasible region is an extreme point of the feasible region (Horst and Tuy, 1996). □

Proof of Theorem 2. The budget constraint is binding in an optimal solution to *RA*. Following the argument in the proof of Theorem 1, INF($\underline{\lambda}$) and $-Q(\underline{\lambda})$ are concave in $\underline{\lambda}$. Let R denote the feasible region of *RA*. The cost function constraint is concave, so R is not necessarily convex. Let R' denote the expansion of R in which the constraint $c_1(\lambda_1) + c_2(\lambda_2) + \cdots + c_M(\lambda_M) \leq B$ has been replaced by its convex hull H. Clearly R' is convex, and every extreme point of R' is in the set R. Let RA' denote the relaxation of *RA* in which R has been replaced by R'. Following the argument in the proof of Theorem 1, the solution to RA' is an extreme point of R'. Since extreme points of R' are also extreme points of R, the optimal solution to *RA* is an extreme point of R. □

Proof of Theorem 3. By Lemma 2 (ii), $NI_i(\lambda_i)$ and $-Q_i(\lambda_i)$ are convex in λ_i for $\lambda_i \in (\max(R_1, a_i), \min(R_2, \delta_i))$ for T sufficiently large. Since positive multiples and sums of convex functions are convex, INF($\underline{\lambda}$) and $-Q(\underline{\lambda})$ are convex in λ_i. INF($\underline{\lambda}$) and $Q(\underline{\lambda})$ are separable in λ_i, so INF($\underline{\lambda}$) and $-Q(\underline{\lambda})$ are convex in $\underline{\lambda}$. Convex cost functions and the linear constraints on $\underline{\lambda}$ lead to a convex feasible region, so over the indicated region of $\underline{\lambda}$, *RA* is a problem of minimizing a convex objective function over a convex feasible region. □

Proof of Theorem 4. The budget constraint is binding in an optimal solution to *RA*. We consider the limit of $NI_i(\lambda_i)$ (given by (7)) as T gets large. For populations with growing epidemics (populations $i = 1, \ldots, m_1$) we have

$$\lim_{T\to\infty} NI_i(\lambda_i) \equiv \frac{\lambda_i - \delta_i}{\lambda_i} - I_{0i} + \delta_i \frac{\lambda_i - \delta_i}{\lambda_i} T \tag{A.10}$$

and for populations with shrinking epidemics (populations $i = m_1 + 1, \ldots, M$) we have

$$\lim_{T\to\infty} NI_i(\lambda_i) \equiv -I_{0i} + \frac{\delta_i}{\lambda_i} \ln\left(\frac{\lambda_i(1 - I_{0i}) - \delta_i}{\lambda_i - \delta_i}\right) \tag{A.11}$$

Consider a solution $\underline{\lambda}'$ with a binding budget constraint in which $\lambda'_j > a_j$ for some population $j \in \{1, \ldots, m_1\}$, and $\lambda'_k < \lambda_{k0}$ for some population $k \in \{m_1 + 1, \ldots, M\}$. There exists another feasible solution $\underline{\lambda}''$, also with a binding budget constraint, that is the same as $\underline{\lambda}'$ except that $\lambda''_k = \lambda_{k0}$ and $\lambda''_j = \lambda'_j - \Delta$, where $\Delta > 0$ satisfies $c_j(\lambda'_j - \Delta) - c_j(\lambda'_j) = c_k(\lambda'_k)$. We have

$$\mathrm{INF}(\underline{\lambda}') - \mathrm{INF}(\underline{\lambda}'') = N_j NI_j(\lambda'_j) - N_j NI_j(\lambda'_j - \Delta) + N_k NI_k(\lambda'_k) - N_k NI_k(\lambda_{k0})$$

It is clear from (A.10) and (A.11) that, for a large enough T, $\text{INF}(\underline{\lambda}') - \text{INF}(\underline{\lambda}'') > 0$, so $\underline{\lambda}'$ cannot be an optimal solution. □

Proof of Theorem 5. Substituting in $q_{i1} - q_{i2} = q$ and $\delta_i = \delta$, we can write (12) as

$$Q(\underline{\lambda}) \equiv \sum_{i=1}^{M} N_i \left[q_{i1}T + \frac{q}{\delta} G_i(\lambda_i) \right] \tag{A.12}$$

Since $q/\delta < 0$, it is clear from (A.12) that for a large enough T, an allocation that maximizes $Q(\underline{\lambda})$ minimizes $\sum_{i=1}^{M} N_i G_i(\lambda_i)$. We now examine $\text{INF}(\underline{\lambda})$, using (7) and (8):

$$\text{INF}(\underline{\lambda}) \equiv \sum_{i=1}^{M} N_i [I_i(T, \lambda_i) - I_{0i} + \delta G_i(\lambda_i)]$$

where

$$I_i(T, \lambda_i) - I_{0i} + \delta G_i(\lambda_i) = \begin{cases} \dfrac{I_{0i}(\lambda_i - \delta_i)\,\mathrm{e}^{(\lambda_i - \delta_i)T}}{\lambda_i I_{0i}(\mathrm{e}^{(\lambda_i - \delta_i)T} - 1) + (\lambda_i - \delta_i)} - I_{0i} & \\ \quad + \dfrac{\delta_i}{\lambda_i} \ln\left[\dfrac{(\lambda_i - \delta_i) + I_{0i}\lambda_i\,\mathrm{e}^{(\lambda_i - \delta_i)T} - 1}{\lambda_i - \delta_i} \right] & \text{for } \lambda_i \neq \delta_i \\ \dfrac{I_{0i}}{\delta_i I_{0i} T + 1} - I_{0i} + \ln(\delta_i I_{0i} T + 1) & \text{for } \lambda_i = \delta_i \end{cases} \tag{8}$$

It is clear from (8) that the term $G_i(\lambda_i)$ dominates for a large enough T. Thus, for a large enough T, an allocation that minimizes $\text{INF}(\underline{\lambda})$ minimizes $\sum_{i=1}^{M} N_i G_i(\lambda_i)$. □

References

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.

Bailey, N.J., 1975. The Mathematical Theory of Infectious Diseases and its Applications. Hafner, New York.

Bernstein, R.S., Sokal, D.C., Seitz, S.T., Auvert, B., Stover, J., Naamara, W., 1998. Simulating the control of a heterosexual HIV epidemic in a severely affected East African city. Interfaces 28, 101–126.

Blount, S., Galamboski, A., Yakowitz, S., 1997. Nonlinear and dynamic programming for epidemic intervention. SIAM Journal of Applied Mathematics and Computation 86, 123–136.

Chen, M.M., Bush, J.W., 1976. Maximizing health system output with political and administrative constraints using mathematical programming. Inquiry 13, 215–227.

Gafni, A., Birch, S., 1993. Guidelines for the adoption of new technologies: a prescription for uncontrolled growth in expenditures and how to avoid the problem. Canadian Medical Association Journal 148, 913–917.

Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A guide to the Theory of NP-Completeness. Freeman, New York.

Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press, New York.

Greenhalgh, D., 1986. Control of an epidemic spreading in a heterogeneously mixing population. Mathematical Biosciences 80, 23–45.

Greenhalgh, D., 1988. Some results on optimal control applied to epidemics. Mathematical Biosciences 88, 125–158.

Hethcote, H.W., 1982. Gonorrhea modeling: a comparison of control methods. Mathematical Biosciences 58, 93–109.

Hethcote, H.W., Van Ark, J.W., 1987. Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. Mathematical Biosciences 84, 85–118.

Hethcote, H.W., Yorke, J.A., 1984. Gonorrhea transmission dynamics and control. Lecture Notes in Biomathematics 56.

Hillier, F.S., Lieberman, G.J., 1993. Operations Research. Holden-Day, New York.

Horst, R., Tuy, H., 1996. Global Optimization: Deterministic Approaches. Springer-Verlag, Berlin.

Kahn, J.G., 1996. The cost-effectiveness of HIV prevention targeting: how much more bang for the buck? American Journal of Public Health 86, 1709–1712.

Kaplan, E.H., 1998. Economic evaluation and HIV prevention community planning: a policy analyst's perspective. In: Holtgrave, D.R. (Ed.), Handbook of Economic Evaluation of HIV Prevention Programs. Plenum Press, New York, pp. 177–193.

Kaplan, E.H., Pollack, H., 1998. Allocating HIV prevention resources. Socio-Economic Planning Sciences 32, 257–263.

Korenromp, E.L., de Vlas, S.J., Nagelkerke, N.J., Habbema, J.D., 2001. Estimating the magnitude of STD cofactor effects on HIV transmission: how well can it be done? Sexually Transmitted Diseases 28, 613–621.

Longini, I.M., Ackerman, E., Elveback, L.R., 1978. An optimization model for influenza A epidemics. Mathematical Biosciences 38, 141–157.

Lopez-Zetina, J., Kerndt, P., Ford, W., Woerhle, T., Weber, M., 2001. Prevalence of HIV and hepatitis B and self-reported injection risk behavior during detention among street-recruited injection drug users in Los Angeles County, 1994–1996. Addiction 96, 589–595.

May, R.M., Anderson, R.M., 1984. Spatial heterogeneity and the design of immunization programs. Mathematical Biosciences 72, 83–111.

Muller, J., 1998. Optimal vaccination patterns in age-structured populations. SIAM Journal on Applied Mathematics 59, 222–241.

ReVelle, C., Male, J., 1970. A mathematical model for determining case finding and treatment activities in tuberculosis control programs. American Review of Respiratory Diseases 102, 403–411.

ReVelle, C., Feldmann, F., Lynn, W., 1969. An optimization model of tuberculosis epidemiology. Management Science 16, B190–B211.

Richter, A., 1996. Optimal resource allocation for epidemic control. Doctoral Dissertation. Operations Research Department, Stanford University, Stanford, CA.

Richter, A., Brandeau, M.L., Owens, D.K., 1999. An analysis of optimal resource allocation for HIV prevention among injection drug users and nonusers. Medical Decision Making 19, 167–179.

Robinson, N.J., Mulder, D.W., Auvert, B., Hayes, R.J., 1995. Modelling the impact of alternative HIV intervention strategies in rural Uganda. AIDS 9, 1263–1270.

Ruiz, M., Gable, A., Kaplan, E.H., Stoto, M., Fineberg, H., Trussell, J. (Eds.), 2001. No Time to Lose: Getting more from HIV Prevention. National Academy Press, Washington, DC.

Sethi, S.P., 1974. Quantitative guidelines for communicable disease control program: a complete synthesis. Biometrics 30, 681–691.

Sethi, S.P., 1978. Optimal quarantine programmes for controlling an epidemic spread. Journal of the Operational Research Society 29, 265–268.

Sethi, S.P., Staats, P.W., 1978. Optimal control of some simple deterministic epidemic models. Journal of the Operational Research Society 29, 129–136.

Stinnett, A.A., Paltiel, A.D., 1996. Mathematical programming for the efficient allocation of health care resources. Journal of Health Economics 15, 641–653.

van Zon, A.H., Kommer, G.J., 1999. Patient flows and optimal health-care resource allocation at the macro-level: a dynamic linear programming approach. Health Care Management Science 2, 87–96.

Varian, H., 1978. Microeconomic Analysis. W. W. Norton, New York.

Weinstein, M.C., 1995. From cost-effectiveness ratios to resource allocation: where to draw the line? In: F.A. Sloan (Ed.), Valuing Health Care: Costs, Benefits, Effectiveness of Pharmaceuticals and other Medical Technologies. Cambridge University Press, New York, pp. 77–97.

Wickwire, K., 1977. Mathematical models for the control of pests and infectious diseases: a survey. Theoretical Population Biology 11, 182–238.

World Health Organization, 1996. WHO information fact sheet no. 110—sexually transmitted diseases. http://www.who.int/inf-fs/en/fact110.html, Accessed 10 January 2002.

World Health Organization, 2000. WHO information fact sheet no. 104—tuberculosis. http://www.who.int/inf-fs/en/fact104.html, Accessed 10 January 2002.

World Health Organization, 2001. AIDS epidemic update—December 2001. http://www.unaids.org/epidemic_update/report_dec01/index.html, Accessed 8 January 2002.

Zaric, G.S., Brandeau, M.L., 2001a. Optimal investment in a portfolio of HIV prevention programs. Medical Decision Making 21, 391–408.

Zaric, G.S., Brandeau, M.L., 2001b. Resource allocation for epidemic control over short time horizons. Mathematical Biosciences 171, 33–58.

Zaric, G.S., Brandeau, M.L., 2003. Dynamic resource allocation for epidemic control in multiple populations. Mathematical Medicine and Biology, in press.