Abstract

Funds spent on HIV prevention commonly traverse several levels of distribution. For example, funds may be allocated to regions, and regional authorities may then allocate their funds to sub-regions or targeted risk groups. Decision makers at each level often make use of heuristics that may result in suboptimal allocation of resources. We examine the impact of equity-based heuristic allocation of HIV prevention funds versus optimal allocation of HIV prevention funds when there are two levels of decision making. Our results demonstrate that if optimization can only be applied to one level of the decision making process, there are more significant gains if it is applied at the lower level than at the upper level.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Health services; Resource allocation; Optimization; HIV/AIDS; Developing countries

1. Introduction

In the last 20 years, over 60 million people have been infected with HIV, and of those cases, 95% are in developing countries [1]. In 2002, the average life expectancy in Sub-Saharan Africa was 47 years, while it would have been 62 years without AIDS [2]. Further, the vast majority of those affected by the disease are in their working years. Over 90% of the world’s HIV-infected children live in sub-Saharan African and mother-to-child transmission (MTCT) is responsible for almost all of those infections [3]. Without treatment, the probability of MTCT during pregnancy, labour and breastfeeding is approximately 30–45% [4].

Investment in HIV prevention prevents loss of life, human suffering, and negative social and developmental consequences. An estimated US$9.2 billion is required annually to implement an effective response to HIV in low- and middle-income countries [5]. This response includes a package of HIV prevention interventions which would avert 29 million infections by 2010 [6]. HIV funding falls short of estimated needs despite considerable recent increases in funding [2]. In 2003, actual spending on HIV/AIDS in resource poor countries totaled approximately $3.6 billion, while the estimated need for HIV/AIDS funding was $6.3 billion [7,8]. Optimization techniques may help to more effectively use
limited resources and narrow the gap between funding needs and availability.

The main sources of external funding for HIV/AIDS interventions in Africa are: donor governments, UN agencies (UNICEF, UNDP, UNESCO, WHO, etc.), the Global Fund to fight AIDS, Tuberculosis and Malaria, and the World Bank [9]. Funds targeted towards HIV interventions traverse several levels of decision making and resource allocation before being spent on the intended interventions. For example, as a centralized funding organization the Global Fund to Fight Aids, Tuberculosis and Malaria begins its financing process with a call for proposals. Proposals are submitted by public and private organizations that collaborate through a Country Coordinating Mechanism (CCM) at a local level. At the highest level of decision making, the Global Fund budget is allocated to countries, via the CCM, based primarily on the scientific merit of their proposal. In turn, the funded countries distribute their funds to local organizations and programs according to their own priorities [10]. Organizations that receive HIV prevention and treatment funds may make further resource allocation decisions before any money is spent on actual prevention and treatment efforts. However, they provide little information about the criteria used to make allocation decisions.

Health economics theory suggests that allocating resources to medical interventions in increasing order of their cost-effectiveness ratios until the available budget has been exhausted will result in the optimal allocation of funds [11]. However, this allocation process does not allow for several important factors such as, increasing or diminishing marginal returns to scale, mutual exclusivity of programs and interaction of program outcomes. Weinstein and Zeckhauser provide a binary integer program [12], and Stinnett and Paltiel develop a mixed integer programming formulation [13] to handle some of these issues. Basso and Peccatti propose a project selection approach which addresses minimum and maximum funding levels and fixed costs using a dynamic programming algorithm [14]. Heidenberger considers the problem of project selection under risk and presents a solution using a stochastic decision tree [15]. Zaric and Brandeau developed a multi-period resource allocation model for epidemic control programs [16].

More comprehensive resource allocation models for control of infectious disease epidemics have been studied. Sanders proposes a control theory approach where the objective is to find the control function which, over time, minimizes the cost of control plus the cost associated with the number of individuals who become infected [17]. Another common approach is to apply simulation in order to compare a set of resource allocation alternatives [18–22].

Dynamic compartmental models are commonly used to model the spread of HIV and population changes over time, and transitions to and from a compartment are typically defined by a system of dynamic equations [21–24]. Brandeau et al. developed a compartmental HIV epidemic model to evaluate the costs and effects of a number of strategies for screening women of childbearing age [25]. Flessa proposes a compartmental systems dynamics model to project the spread of HIV, and then assesses the impact of health education, prevention of mother-to-child-transmission and hypothetical vaccination programs using simulation [26]. Similar approaches have also been suggested for malaria control [27].

A number of approaches have been developed in which the epidemic control problem is formulated as a nonlinear optimization problem. The problem is usually stated as one of choosing the amount to be invested in several interventions to optimize total health benefits subject to a budget constraint [28]. Linear programming models of resource allocations for health care in general have been developed; they suggest that the outcome measure used in the objective function has a strong influence on the resulting optimal allocation [29,30]. Kaplan, and Kaplan and Pollack propose a nonlinear optimization formulation aimed at maximizing the number of averted HIV infections which uses a dynamic programming method for determining production functions [31,32]. Zaric and Brandeau developed approximations for allocating epidemic control resources over short time horizons [33]. Nonlinear optimization approaches are summarized by Zaric [28].

Funds targeted towards HIV interventions are commonly allocated based on equity criteria, such as proportional to the number of HIV/AIDS cases in different subgroups [34,35]. Zaric and Brandeau showed the impact of using equity heuristics [36], and Kaplan and Pollack showed what assumptions need to be made in order for equity based heuristics to be optimal [32]. HIV policy models have been suggested, but most relate to developed countries, and few models are specifically based on developing countries [37–39]. Conclusions drawn from studies in developed country settings are not generalizable
to developing countries due to differences in the causes of epidemic proliferation, the cost-effectiveness of interventions and available resources.

In this paper, we evaluate the impact of simple resource allocation heuristics versus optimal allocation, given two levels of decision-making, in the context of Sub-Saharan Africa. If additional effort and rigour, such as that involved in optimization modelling, can only be applied to the decision making process at one level, we determine whether it should be at the higher or lower level of decision making. This research is innovative in its inclusion of both sexual contact and MTCT as modes of HIV transmission and in its evaluation of the impact of more than one level of allocation decision. Our results are intended to help policy makers in governments, public health agencies and non-government organizations make informed decisions regarding AIDS policy modelling and budget allocation. The remainder of this paper is organized as follows: we begin with a description of the mathematical models used; we then describe the results of our baseline and sensitivity analysis scenarios. Finally, we conclude with an interpretation of our results, some known limitations and suggestions for future work.

2. Methods

We model a two-level resource allocation process, as depicted in Fig. 1. We refer to the structure depicted in Fig. 1 as an “allocation network”. We assume a one-time allocation of funds to populations within this network over a fixed period. HIV prevalence rates tend to be higher among specific groups, such as commercial sex workers or young women attending antenatal clinics in urban areas. Thus, the allocation network is composed of two regions, Region 1 and Region 2, each of which is divided into two subpopulations which we refer to as high-risk and low-risk. Within each subpopulation, resources are allocated towards two general types of interventions: reducing HIV transmission rates resulting from unsafe sexual contact, and reducing MTCT, also known as vertical transmission. The upper level allocation decision consists of determining the total funds that should be invested in Region 1 and Region 2. The lower level allocation decision, which is made at the regional level, consists of determining the amounts that should be invested in intervention programs aimed at reducing the rates of unsafe sexual contact and MTCT, within each of the subpopulations.

Optimality is defined as an allocation strategy for which the results cannot be improved upon while equity implies an allocation strategy based on proportionality or perceived fairness. For this analysis, we define equity allocation as proportional to the number of individuals infected with HIV.

We developed four methods of solving the resource allocation problem, which we refer to as Optimal–Optimal, Optimal–Equity, Equity–Optimal, and Equity–Equity. In Optimal–Optimal, we substitute the lower level problem into the upper level problem and optimally solve the resulting allocation problem for all decision variables. Thus, the Optimal–Optimal solution method looks at the problem as if there were a single level of decision making. The solution to this method represents a lower bound on the number of new infections that can be achieved by any solution method. In Optimal–Equity, we bind the upper level problem to the proportions determined by the equity solution in the lower level problem. In Equity–Optimal, we
use the equity model to make the upper level allocation, and then independently solve the two lower level problems optimally. In Equity–Equity we solve the upper and lower level problems using an equity heuristic.

2.1. Epidemic model

We develop an epidemic model to determine the outcome of different allocation decisions. We define the epidemic model as a nonlinear system of differential equations. The formulation of the set of ordinary differential equations draws on the structure of an S-I (Susceptible-Infected) model with sufficient contact rate, \( \lambda \), vertical transmission rate, \( m \), and varying population size. Compartmental models like ours are commonly used for defining an epidemic model and have been used extensively elsewhere [24,40–43].

Our model is illustrated in Fig. 2. Each of the four subpopulations is modelled as a four-compart-ment epidemic model where the population is divided into two disease states (infected or not) and sexual activity states (active or not). Let \( S(t) \), \( I(t) \), \( U(t) \) and \( V(t) \) represent the numbers of susceptible adults, infected adults, uninfected children and infected children, respectively, at time \( t \). Members of the infected children compartment, \( V(t) \), are those who were infected by vertical transmission and are not in the pool of sexually active adults. We assume that members of \( V(t) \) and \( U(t) \) are not sexually active and thus cannot contract or spread the virus.

Let \( i \) be an index over the subpopulation (high or low-risk) and \( j \) be an index over the region (\( i = 1,2; j = 1,2 \)). Since the model is the same in each subgroup we suppress \( i, j \) from the notation except where necessary. The model in subgroup \( i, j \) is specified as follows:

\[
S'(t) = -\lambda \cdot I(t)S(t) - \delta_S \cdot S(t) + \gamma \cdot U(t), \quad (1)
\]

\[
I'(t) = \lambda \cdot I(t)S(t) - \delta_I \cdot I(t), \quad (2)
\]

\[
U'(t) = -\delta_U \cdot U(t) + \beta \cdot S(t)
+ (1 - m) \cdot \beta \cdot I(t) - \gamma \cdot U(t), \quad (3)
\]

\[
V'(t) = -\delta_V \cdot V(t) + m \cdot \beta \cdot I(t). \quad (4)
\]

In Eqs. (1)–(4), the death rates for the susceptible adults, infected adults, uninfected children and infected children compartments are noted by, \( \delta_S \), \( \delta_I \), \( \delta_U \), and \( \delta_V \) respectively. The birth rate is denoted by \( \beta \). The probability of vertical transmission, which includes the probabilities of transmission during pregnancy, labour, delivery and through breastfeeding, is denoted by \( m \).

---

Fig. 2. Illustration of a four compartment epidemic model.
The sufficient contact rate is denoted by \( \lambda \). A sufficient contact is a sexual encounter that results in infection of a susceptible individual by an infected individual. We aggregate adult modes of transmission over all types of sexual contacts, including male-to-female and female-to-male heterosexual, as well as homosexual activities. We aggregate for three reasons. First, data on mode of transmission in Africa is scarce. Second, the lifetime probability of contracting HIV is approximately equal for men and women. Finally, while not the chief mode of transmission in Africa, homosexual contact is still a significant mode of transmission [44–49]. Because homosexual contact is illegal in many countries in Africa, it may be under-reported in official statistics. However, one source claims that it accounted for approximately 8% of all adult HIV transmission in South Africa [50]. Thus, we chose to use a model with aggregate compartments and aggregate sufficient contact rates that could be estimated directly from reported data rather than to design additional models to allow us to produce estimates of disaggregated values.

Maturation is achieved by transferring a portion of the uninfected children subpopulation to its corresponding adult subpopulation while considering non-AIDS related deaths. The maturation rate is denoted by \( \gamma \). Given the very high mortality rate among infected children, we estimate that only 4.3% of that subpopulation would survive to the age of 15 [51,52]. We assume that members of that subpopulation do not enter the sexually active infected adult population. The infected children compartment is used to count pediatric infections and life years.

Our model is based on 2001 population and HIV/AIDS prevalence rates of Kenya and Botswana established by the Population Division of the United Nations Secretariat [53]. We validate our epidemic projections by fitting the model to reproduce the historical behavior in Kenya and Botswana between the years 1995 and 2001.

### 2.2. Production functions

Economists typically define a production function, \( f(v) \), as the level of output that can be achieved from input amounts \( v = (v_1, v_2, \ldots, v_n) \) [54]. In the context of HIV epidemic control, production functions are intended to convert “dollar amounts invested” into changes in the sufficient contact rate (\( \lambda \)) and vertical transmission rate (\( m \)). They help to derive the number of new infections that occur in a particular population for a given level of investment.

There are several ways to define production functions. In their simplest form, production functions can be expressed as a linear relationship between the investment in a prevention program and the reduction in the rate of HIV transmission. Complexities such as fixed start-up costs and increasing or diminishing returns to scale can be introduced [55]. Intervention programs are typically bound by a maximum level of effectiveness; this is a lower limit beyond which additional investments will not yield further results. For example, there may be marginalized communities that will never be reached by a given intervention, or some individuals will be resilient to certain types of behavioral change.

We model an exponential production function for HIV prevention programs; the exponential form implies diminishing returns to scale. Let \( x \) be the amount invested in interventions aimed at reducing the contact rate. The corresponding production function is expressed by

\[
\lambda_{\text{final}} = \lambda_{\text{initial}} \cdot (a - (a - 1) \cdot e^{-bx}).
\]  

(5)

When \( x \) tends towards infinity, the maximum level of effectiveness is attained and \( \lambda_{\text{final}} = \lambda_{\text{initial}} \cdot a \). Thus, \( \lambda_{\text{final}} \) is reduced to a fraction of \( \lambda_{\text{initial}} \), and this fraction is represented by parameter \( a \), which is a multiplier between 0 and 1. Parameter \( b \) represents a negative change in the sufficient contact rate per dollar invested and is derived from the cost-effectiveness ratios of programs that reduce sexual transmission of HIV. Details are provided in Appendix A.

A similar production function is applied to \( m \), the rate of MTCT where \( y \) is the amount invested in interventions aimed at reducing MTCT, \( p \) is a multiplier that reduces the initial MTCT rate to its lower limit and \( q \) is the change in the rate of MTCT per dollar invested in interventions aimed at reducing MTCT. The corresponding production function is expressed by

\[
m_{\text{final}} = m_{\text{initial}} \cdot (p - (p - 1) \cdot e^{-qy}).
\]  

(6)

In our model we index all variables in the production functions by \( i, j \) where \( i \) is an index over the subpopulation and \( j \) is an index over the region \( (i = 1, 2; j = 1, 2) \).
2.3. Optimization model

One method of allocating resources at each level of decision making is to formally specify an optimization model. Consistent with other work on HIV prevention, we assume that the objective is to minimize the number of new infections given a total budget constraint (e.g., [32]). We are not aware of a closed-form analytic solution for the epidemic model specified by (1)–(4). Thus, we estimate the cumulative number of new infections between time zero and time $T$ using discrete-time approximations to the continuous system in short time intervals [56,57]. The cumulative number of new infections in region $j$, noted $N_j$, is expressed as follows:

$$
N_j(T) = \sum_i \int_0^T \left[ \alpha_{ij} \cdot I_{ij}(t)S_j(t) + m_{ij} \cdot \beta_{ij} \cdot I_{ij}(t) \right] dt,
$$

$i = 1, 2.$

(7)

The first term of the integrand represents new adult infections in period $t$ while the second term represents new infant infections in period $t$.

In our optimization model we minimize the cumulative number of new infections by reducing $\lambda$ and $m$, which are a function of their allocated budgets. Let $x_{ij}$ be the amount invested in interventions aimed at reducing the contact rate in subpopulation $i$ of region $j$. Let $y_{ij}$ be the amount invested in interventions aimed at reducing MTCT in subpopulation $i$ of region $j$. The lower level allocation model for a specific region $j$ is expressed as follows:

$$
\text{Min}_{x_{ij},y_{ij}} \sum_i N_j(x_{ij},y_{ij}),
$$

subject to:

$$
x_{ij}, y_{ij} \geq 0 \quad i = 1, 2.
$$

(8)

(9)

The decision variable $x$ represents the amount allocated towards reducing $\lambda$ and $y$ represents the amount allocated towards reducing $m$. $B_j$ is the budget allocated to region $j$. The production functions defined in (6) and (7) are used to establish $\lambda$ and $m$; they are a function of the amounts invested in interventions aimed at reducing unsafe sexual contact and reducing vertical transmission, respectively. The invested amounts are the decision variables of the optimization model, which are constrained by a maximum budget, in (9), and by non-negativity, in (10).

The upper level allocation model is expressed as follows:

$$
\text{Min}_{B_1,B_2} N_1(B_1) + N_2(B_2)
$$

subject to:

$$
B_1 + B_2 \leq B,
$$

$$
B_1, B_2 \geq 0,
$$

(11)

(12)

(13)

where the decision variables, $B_1$ and $B_2$, represent the amounts allocated to Regions 1 and 2, respectively, and $B$ is the total budget available to Regions 1 and 2. The description of model (11)–(13) is analogous to the description of model (8)–(10); however, the upper level model allocates the budget to both Regions 1 and 2.

2.4. Equity model

Resources at each level may be allocated on the basis of equity. We define equity as proportional to the number of HIV cases in each region. Under this method, budget $B$ is allocated to Regions 1 and 2 proportionally, based on their share of the total infected population. The regional budget, $B_j$, is then allocated towards $\lambda$ and $m$ of the high-risk and low-risk subpopulations proportionally, based on the number of HIV cases in the adult and children compartments of the subpopulations.

The upper level allocation model is expressed as follows:

$$
B_j = B \times \frac{I_j(0) + V_j(0)}{\sum_j(I_j(0) + V_j(0))};
$$

(14)

the lower level allocation model for region $j$ is expressed as follows:

$$
B_{ij} = B_j \times \frac{I_{ij}(0) + V_{ij}(0)}{\sum_i(I_{ij}(0) + V_{ij}(0))}, \quad i = 1, 2;
$$

(15)

$$
B_{2ij} = B_{ij} \times \frac{I_{ij}(0)}{I_{ij}(0) + V_{ij}(0)}, \quad i = 1, 2;
$$

(16)

$$
B_{mij} = B_{ij} \times \frac{V_{ij}(0)}{I_{ij}(0) + V_{ij}(0)}, \quad i = 1, 2.
$$

(17)

$B_j$ represents the budget allocated to region $j$, $B_{ij}$ represents the budget allocated to subpopulation $i$ of region $j$, $B_{2ij}$ represents the budget allocated to reducing the contact rate in subpopulation $i$ of region $j$ and $B_{mij}$ represents the budget allocated to reducing MTCT in subpopulation $i$ of region $j$. A number of similar equity measures have been discussed elsewhere [32].

2.5. Estimation of parameter values

We modelled the populations of Regions 1 and 2 based on the 2001 year-end population and HIV/
AIDS prevalence rates of Kenya and Botswana. We define Region 1 as a hypothetical cohort of 31.3 million individuals and Region 2 as a hypothetical cohort of 1.5 million individuals. Parameter baseline values and their sources are listed in Table 1. The low-risk subpopulations within the regions were derived based on the percentage of the population living in rural areas, while the high-risk subpopulations were derived based on the percentage of the population living in urban areas. Prevalence in the adult population (aged 15–49) of Region 1 is 17% and 10% for the high and low-risk subpopulations; these rates are 40% and 26% for Region 2, respectively. These estimates are supported by the higher HIV prevalence rates in woman attending antenatal clinics in urban areas [2]. Table 2 shows the number of people in each subpopulation by compartment. In our baseline scenario, we assumed a budget of $40 million dollars, reflecting approximately the funds committed by the Global Fund to Kenya and Botswana in the first two rounds of funding. We also assumed a time horizon of 20 years to capture the dynamics of the epidemic in our results.

The WHO sub-region AFRO-E consists of 20 countries in Africa that WHO defines as having “very high” adult mortality and “high” child mortality [58]. The birth rate, $\beta$, was derived from the

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Baseline values</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth rate</td>
<td>$\beta = 3.65%$</td>
<td>Global Burden of Disease [58]</td>
</tr>
<tr>
<td>Death rates</td>
<td>$\delta_S = 1.2%$; $\delta_I = 7.5%$; $\delta_U = 1.5%$ and $\delta_V = 18.9%$</td>
<td>Report on the Global HIV/AIDS Epidemic [2]; Global Burden of Disease 2000 [58]</td>
</tr>
<tr>
<td>Initial sufficient contact rate ($\lambda_{initial}$)</td>
<td>High-risk subpopulation: 16%; Low-risk subpopulation: 10%</td>
<td>Report on the Global HIV/AIDS Epidemic [2]; Global Burden of Disease 2000 [58]</td>
</tr>
<tr>
<td>Initial MTCT rate ($m_{initial}$)</td>
<td>High-risk subpopulation: 31%; Low-risk subpopulation: 19%</td>
<td>De Cock et al. (2000) [4]</td>
</tr>
<tr>
<td>Cost-effectiveness ratio of interventions aimed at reducing unsafe sexual contact</td>
<td>$80\text{S per infection averted}$</td>
<td>Creese et al. (2002) [60]; Walker [61]</td>
</tr>
<tr>
<td>Cost-effectiveness ratio of interventions aimed at reducing MTCT</td>
<td>$175\text{S per infection averted}$</td>
<td>Creese et al. [60]; Rauner et al. [19]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population</th>
<th>Region 1: 31.3 million</th>
<th>Region 2: 1.5 million</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of population living in urban areas</td>
<td>Region 1: 34%; Region 2: 49%</td>
<td>World Population Prospects [53]</td>
</tr>
</tbody>
</table>

Table 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Region 1</th>
<th>Region 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-risk</td>
<td>High-risk</td>
<td>Low-risk</td>
</tr>
<tr>
<td>$S(0)$</td>
<td>10,500,000</td>
<td>5,000,000</td>
</tr>
<tr>
<td>$H(0)$</td>
<td>1,200,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>$U(0)$</td>
<td>8,900,000</td>
<td>4,500,000</td>
</tr>
<tr>
<td>$V(0)$</td>
<td>110,000</td>
<td>90,000</td>
</tr>
<tr>
<td>Totals</td>
<td>20,710,000</td>
<td>10,590,000</td>
</tr>
</tbody>
</table>

Table 2

b The World Bank Group, World Development Indicators database, April 2003.

average number of children born to women between the ages of 15 and 44 in the AFRO-E sub-region. The same birth rate was used for all regions, subpopulations and disease states. Death rates $\delta_S$, $\delta_I$, $\delta_U$, and $\delta_V$ were derived from the average death rate by states of disease and sexual activity in the AFRO-E sub-region. The age of sexual maturity, $\tau$, was defined as 15 based on a lower bound of the average median age of first sexual contact in the countries of the AFRO-E sub-region [2]. In the uninfected children compartment, we estimate the number of persons alive, for each distinct age between 0 and 14, based on the annual mortality and population growth rates. We then divide the number of 14 year olds by the total size of the group. This results in a rate which represents the
number of children who will turn 15 and thus enter
the compartment of sexually active uninfected
adults. This rate is \( r \), the maturation rate.

The actual number of new HIV infections in
adults can be approximated by \( k \cdot [I(t) \cdot S(t)] \). We
use the number of new HIV infections in adults,
denoted by \( \Psi \times [I(0) + S(0)] \), in the AFRO-E sub-
region, for the year 2000 [2,58,59] to estimate the
initial sufficient contact rate, as follows:

\[
\lambda_{\text{initial}} = \frac{\Psi}{[I(0) \cdot S(0)]} \times [I(0) + S(0)].
\]  

(18)

The initial MTCT rate, \( m_{\text{initial}} \), was defined as
25% based on reported estimates [4]. We assumed
that \( \lambda \) and \( m \) varied by risk group but were the same
across regions for comparable risk groups. We esti-

Table 3
Results of baseline scenario

<table>
<thead>
<tr>
<th>Solution method</th>
<th>Region</th>
<th>Region 1</th>
<th>Region 2</th>
<th>Total</th>
<th>Region</th>
<th>Region 1</th>
<th>Region 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal–Optimal</td>
<td>Region 1</td>
<td>2,143,041</td>
<td>234,038</td>
<td>2,377,079</td>
<td>Region 1</td>
<td>2,392,733</td>
<td>178,941</td>
<td>2,571,674</td>
</tr>
<tr>
<td></td>
<td>Region 2</td>
<td>36,550,000</td>
<td>3,450,000</td>
<td>40,000,000</td>
<td>Region 2</td>
<td>35,200,000</td>
<td>4,800,000</td>
<td>40,000,000</td>
</tr>
<tr>
<td>Rank</td>
<td>1st</td>
<td></td>
<td></td>
<td></td>
<td>4th</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Optimality–Equity ranked second with a total of 2,392,557 new infections and Optimal–Equity ranked third with a total of 2,552,393 new infections. These results

These interventions yield an average cost of $80 per averted infection which results in an estimate of \( b = 5.14 \times 10^{-8} \) (see Appendix A for details). To estimate \( q \), we assessed the average cost-effectiveness ratio in dollars per averted vertical transmission for prevention programs aimed at reducing MTCT in sub-Saharan Africa. Interventions aimed at reducing

The model was formulated in Microsoft Excel. Optimization was performed using the Premium Solver Platform version 5.1 from Frontline Systems. We validated the optimality of our solver results by comparing them to the results of grid searches over the feasible region for each decision variable. We used one year time increments in the discrete time approximation of the epidemic model.

3. Results

We compared the four methods for allocating resources at two levels of decision making. Our results are summarized in Table 3. In the absence of any investment in HIV prevention, there would be 7,577,097 new infections over 20 years. We found that Optimal–Optimal is the method that minimizes the total number of new infections with 2,377,079 new infections. This method calls for 91.4% of the budget to be allocated to Region 1 and the remaining 8.6% of the budget allocated to Region 2. Equity–Equity ranks last with 2,571,674 new infections, and 88% and 12% of the budget allocated to Regions 1 and 2, respectively. Equity–Optimal ranked second with a total of 2,392,557 new infections and Optimal–Equity ranked third with a total of 2,552,393 new infections. These results
suggest that the lower level allocation is of greater consequence in terms of minimizing the number of new infections than the upper level allocation.

The Optimal–Equity solution method yields an improvement of less than 1% compared to Equity–Equity, the Equity–Optimal method yields an improvement of approximately 7% over the outcome of Optimal–Equity, and the Optimal–Optimal method yields an improvement of less than 1% over the outcome of Equity–Optimal. The gap between best and worst solution method is 194,595 new infections, representing an overall 8% increase in the number of new infections.

3.1. Sensitivity analysis

We performed univariate sensitivity analysis on the budget, the time horizon, the prevalence rates, the cost-effectiveness of interventions for reducing unsafe sexual contact and MTCT and the maximum reduction in the rates of unsafe sexual contact and MTCT. Table 4 is a summary of the sensitivity analysis; it describes the ranges used as well as the gap between the Optimal–Optimal and Equity–Equity solution methods for the upper and lower bounds of the sensitivity analysis ranges. If the gap is wide, this demonstrates that the solution method used has a considerable impact on health outcomes; while if the gap is narrow this demonstrates that the solution methodology does not significantly influence the overall results.

As the budget is varied between $20 million and $400 million, the gap narrows from 6% to 1%. The time horizon was adjusted from 20 years to 5, 10, 50 and 100 years. The gap widens as the time horizon increases because the effects of the original misallocation are accrued over a larger period of time. The prevalence rates in all subpopulations were varied by substituting a percentage of the non-infected populations (S and U) to the infected populations (I and V). As a result, the prevalence is Region 1 was varied from 7% to 20% and from 28% to 42% in Region 2. The increase in prevalence puts more strain on the fixed budget, so the gaps widens accordingly. Also, as the overall prevalence rate increases so does the allocation to interventions aimed at reducing MTCT.

In the baseline scenario, averting an infection transmitted from mother-to-child is more than twice as expensive as averting an infection transmitted through unsafe sexual contact. So, given the populations we consider, there are significantly more funds allocated towards reducing unsafe sexual contact than reducing MTCT. The cost-effectiveness of interventions for reducing unsafe sexual contact was varied from $40 to $800 per infection averted while the cost-effectiveness of interventions for reducing MTCT was varied from $85 to $1750 per infection averted [60,61]. However, even when varying the cost-effectiveness of interventions for reducing unsafe sexual contact and MTCT, to the advantage of MTCT, there remains a significantly larger portion of the budget allocated towards reducing unsafe sexual contact.

Parameters \( a \) and \( p \) were varied simultaneously between 0% and 50%. Multiplied by \( \lambda_{\text{initial}} \) and \( m_{\text{initial}} \), these parameters represent the lower bound on \( \lambda_{\text{final}} \) and \( m_{\text{final}} \). At 0%, unsafe sexual contact and MTCT can be completely eliminated; at 50%, unsafe sexual contact and MTCT can at best be

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Baseline value</th>
<th>Lower Bound(^a)</th>
<th>Gap Lower(^b) (%)</th>
<th>Upper Bound(^a)</th>
<th>Gap Upper(^c) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost-effectiveness of interventions for reducing unsafe sexual contact ($ per averted infection)</td>
<td>80</td>
<td>40</td>
<td>7</td>
<td>800</td>
<td>3</td>
</tr>
<tr>
<td>Cost-effectiveness of interventions for reducing MTCT($ per averted infection)</td>
<td>175</td>
<td>18</td>
<td>6</td>
<td>1750</td>
<td>13</td>
</tr>
<tr>
<td>Parameters ( a ) and ( p ), i.e. maximum reduction in ( \lambda_{\text{final}} ) and ( m_{\text{final}} ) as a fraction of ( \lambda_{\text{initial}} ) and ( m_{\text{initial}} )</td>
<td>0.1</td>
<td>0</td>
<td>9</td>
<td>0.5</td>
<td>4</td>
</tr>
<tr>
<td>Total Budget (million $)</td>
<td>40</td>
<td>20</td>
<td>6</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>Time Horizon (years)</td>
<td>20</td>
<td>5</td>
<td>3</td>
<td>100</td>
<td>52</td>
</tr>
<tr>
<td>Prevalence in Region 1 (R1) &amp; Region (R2)</td>
<td>( R1 : 8% ) ( R1 : 7% )</td>
<td>( R1 : 28% )</td>
<td>( R2 : 21% ) ( R2 : 20% )</td>
<td>( R2 : 42% )</td>
<td>21</td>
</tr>
<tr>
<td>Weight for infant infections</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>50</td>
<td>37</td>
</tr>
</tbody>
</table>

\(^a\) Lower bound and Upper Bound represent the minimum and maximum parameters values used in sensitivity analysis.

\(^b\) Gap Lower represents the difference between the Optimal–Optimal and the Equity–Equity solution methods at the lower bound.

\(^c\) Gap Upper represents the difference between the Optimal–Optimal and the Equity–Equity solution methods at the upper bound.
halved. For every value of $a$ and $p$, ranking of all solution methods is maintained as per the base case. Also, the number of new infection increases as $a$ and $p$ are increased confirming that the lower bounds on $\lambda_{\text{final}}$ and $m_{\text{final}}$ are constraining.

We evaluated more than 200 scenarios in sensitivity analysis. The ranking of the four options was maintained as in the base case. Equity–Optimal was preferred over Optimal–Equity option in all cases.

3.1.1. Allocation to adult vs. pediatric infections

In all scenarios evaluated, the optimal investment called for more resources to be devoted to reducing unsafe sexual contact than to reducing vertical transmission. In the baseline case, the Optimal–Optimal solution method results in an allocation to the prevention of MTCT representing only 0.1% of the total budget. Considering that infected infants will not live to potentially transmit the virus through unsafe sexual contact, it is theoretically less effective to avert MTCT. The baseline case considers all infections averted to be weighted equally, regardless of the mode of transmission.

In this scenario of sensitivity analysis, we added a weighting factor to the new infant infections since the number of life years lost in children is greater than that of adults when ignoring the life years lost from secondary infections. We assume that an averted pediatric HIV infection results in three times the number of life years gained compared to an averted adult HIV infection. Therefore, we initially considered a weighting factor of 3; this did not change the ranking of the four solution methods. Using the Optimal–Optimal solution method, the total number of new infections was increased by 1% while the number of pediatric infection was reduced by 48%; and allocation to the prevention of MTCT increased to 13% of the total budget. A weighting factor of at least 50 would be required before there was equal spending on both modes of transmission.

3.1.2. Birth rate: A function of investment in interventions

There are many biological, behavioral, socio-economical and cultural factors that influence fertility changes in the context of HIV [62,63]. It is widely accepted that HIV exerts a downward pressure on overall fertility [62]. Lewis et al. [64] conducted a review of 19 studies to assess the impact of HIV on fertility in sub-Saharan Africa. According to this study, the relationship between HIV prevalence and its population level impact follows a linear relationship. We use a function similar to the production function to model the impact of HIV prevention interventions aimed at reducing unsafe sexual contact onto the per capita birth rate. This linear function is expressed as follows:

$$\beta_{\text{final},ij} = \beta_{\text{initial},ij} - \sum_{ij} (e \cdot x_{ij}), \quad (19)$$

$\beta_{\text{initial}}$ is the birth rate used in our baseline scenario and $\beta_{\text{final}}$ is the birth rate used for the sensitivity analysis scenarios. The decision variable $x$ represents the amount allocated towards reducing unsafe sexual contact. Parameter $e$ represents the influence of investing in interventions aimed at reducing unsafe sexual contact onto the birth rate, including condom distribution. Parameter $e$ was varied to mimic an overall decline in total fertility between 5% and 50%. This range was selected based on the average difference between the total fertility rate and the total wanted fertility rate in Africa which stands at 27% [65]. Results of this scenario indicate that though the number of new infections averted is decreased by virtue of the smaller number of births, the ranking of the four options was maintained as in the base case.

3.1.3. Production functions

We assessed the outcome of using linear production functions for determining $\lambda$ and $m$. We use a linear production function of the form $\varphi_{\text{final}} = \varphi_{\text{initial}} + (c \cdot x)$ where $x$ is amount invested in interventions aimed at reducing the contact rate, and $c$ is the slope representing the change in $\varphi_{\text{initial}}$ per amount invested. Given the cost-effectiveness ratio of investing in interventions aimed at reducing unsafe sexual contact, we identify $c$ by fitting the linear curve between two known points. We use an analogous linear function to determine $m$, the rate of MTCT. Results of this scenario reveal the same ranking of the four solution methods as the base case scenario. The gap when using the baseline exponential production function for determining $\lambda$ and $m$ is 8%, while this gap widens to 94% when using a linear production function for determining $\lambda$ and $m$. This stresses the importance of production functions on the overall outcome. We believe that exponential functions are more suitable than linear functions for large extrapolations, because exponential functions are flatter and show diminishing returns for large investments thereby narrowing the gap.
between the best and worst solution methodology. Brandeau et al. have found similar results [55].

An initial investment of funds may be required for interventions to demonstrate any positive effect. Therefore, we created a production function which includes a start-up cost below which investment in prevention interventions has no effect and above which the investments follow the exponential production function stated in Eq. (5). This discontinuous production function is expressed as

\[
\hat{\lambda}_{\text{final}} = \begin{cases} 
\hat{\lambda}_{\text{initial}} \cdot (a - (a - 1) \cdot e^{-bx} & \text{if } x \geq \text{start-up costs}, \\
\hat{\lambda}_{\text{initial}} & \text{otherwise.}
\end{cases}
\]

We maintain the total available budget at $40 million and sequentially set the start-up costs for the two types of intervention programs in each of the four compartments to zero, $1 million, and then $2 million. Since the start-up costs are additional constraints to the model, the number of new infections increases with the start-up cost. The Optimal–Equity option bears 0.5% more infections with a start-up cost of $2 million than with no start-up costs and the Equity–Equity option bears 9.6% more infections with a start-up cost of $2 million than with no start-up costs. In terms of new infections, the gap between the best and worst solution method is 8% with no start-up costs, and 9% and 17% with start-up costs of $1 million and $2 million, respectively. In this scenario, the ranking of the four options was maintained as per the base case.

4. Discussion

In this paper we examined allocation of HIV prevention funds when there are two levels of decision making. In our baseline scenario, the number of new infections in the Equity–Optimal option represents an improvement of 7% over the number of new infections in the Optimal–Equity option. We conclude that if optimization can only be applied to one level of the decision making process, there are more significant gains if it is applied at the lower level than at the upper level. This conclusion has several implications. Our analysis suggest that focussing solely on higher level allocations can yield ineffective use of scarce resources, and that knowledge of lower level data is crucial and attention should be brought to the manner in which funds are allocated towards lower levels.

The robustness of the model is established through extensive sensitivity analyses. We found that the ranking of the four options evaluated is not sensitive to variations in six of the model’s parameters. Though we cannot analytically demonstrate the superiority of the Equity–Optimal option over the Optimal–Equity option, the advantage of the Equity–Optimal option was maintained across all sensitivity analysis scenarios considered.

Our results must be viewed in the context of several limitations. First, though our epidemic model can accommodate multiple time periods, the optimization model is modelled as a single period. This means that budgets allocated to the subpopulations will be entirely consumed at the start of the time horizon rather than being spent over a series of time periods. As well, the impact of the funds allocated on the transmission rates, which is calculated through the production functions, will be immediate and permanent. Techniques to transform the optimization model into a multi-period model are known [16]. This would enable consideration of multiple budgets over time and the effects of the investments could taper off if they are not maintained.

This study should be viewed as a first step towards further evaluation of resource allocation models for epidemic control given several levels of allocation decisions. In addition, production functions can be improved upon by including: a distinct breakdown of interventions, minimum or maximum program funding levels, increasing or diminishing marginal returns, mutual exclusivity of programs and interaction of program outcomes. Further, anti-retroviral therapy (ART) is currently being introduced in several countries of sub-Saharan Africa. The impact of ART could be added to our epidemic transmission model once studies provide an understanding of the net impact of this influence.

Our results and conclusions hold true for a model consisting of two regions, eight subpopulations and two levels of allocations decisions. It is of interest to broaden this research in an attempt generalize the conclusions to \(k\) regions, \(m\) subpopulations and \(n\) levels of allocations decisions. The approach suggested in this paper can be used to compare other resource allocation methods, for example, more sophisticated equity allocation models or numerical analysis techniques such as simulation.

The causes of HIV proliferation, the parameterization of the model and the assumed resource scarcity were selected to represent the context of sub-Saharan Africa. Though we cannot generalize
the conclusions drawn from this model to other settings, the model and solution procedures as a whole are likely to be applicable to other diseases and in other geographic locations.

AIDS seriously weakens the economies and healthcare systems of developing countries. New data show that the epidemic has not reached a natural limit in Africa, and HIV prevalence continues to rise [2]. This study demonstrates that resource allocation methods for HIV prevention have a strong influence on the overall outcome of the disease and that serious consideration should be given to optimizing the lower levels of decision-making.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by an Ontario Graduate Scholarship in Science and Technology (OGSST). GSZ also received support from the National Institute on Drug Abuse (DA-R01-15612).

Appendix A. Finding the exponential production function parameters

The production function is defined as follows:

$$\lambda_{\text{final}} = \lambda_{\text{initial}} \cdot (a - d \cdot e^{-bx})$$

where parameters $a$, $d$, and $b$ need to be identified and the second term in the function, $f(x) = a - d \cdot e^{-bx}$, is a multiplier, so $f(x) \in [0, 1]$.

A.1. Identifying parameter $a$

We assume that, at best, $\lambda_{\text{initial}}$ can be reduced to a fraction of its initial value; let $g$ represent this fraction. Investing an infinite amount of funds will yield the maximum benefit and the maximum reduction of $\lambda_{\text{initial}}$ will be attained, and $\lambda_{\text{final}} = \lambda_{\text{initial}} \cdot g$. So,

$$a - d \cdot e^{-bx} = g,$$
$$d \cdot e^{-bx} = a - g,$$
$$d \cdot e^{-bx} = a - g,$$
$$0 = a - g,$$
$$g = a.$$

A.2. Identifying parameter $d$

Investing no funds would have no impact on $\lambda_{\text{final}}$, thus the multiplier would be 1. In this case, $\lambda_{\text{final}} = \lambda_{\text{initial}}$. So,

$$a - d \cdot e^{-bx} = 1,$$
$$a - d \cdot e^{bx} = 1,$$
$$a - d = 1,$$
$$d = a - 1.$$

A.3. Identifying parameter $b$

For each region, we use our model at time zero to estimate the number of new infections, after one year, resulting from unsafe sexual contact if no funds are invested. We then determine the corresponding sufficient contact rate for sexual contacts, $\lambda_{\text{initial}}$, using Eq. (18). Given that the average cost-effectiveness of prevention programs aimed at reducing unsafe sexual contact is $80 per averted infection, we then establish $\theta$, the rate of reduction in the sexual contact rate resulting from an investment of $x$ equal to $80$. In this case, an investment of $80 yields $\lambda_{\text{final}} = \lambda_{\text{initial}} \cdot (1 - \theta)$. So,

$$a - d \cdot e^{-bx} = 1 - \theta,$$
$$-d \cdot e^{-bx} = (1 - \theta) - a,$$
$$d \cdot e^{-bx} = a - (1 - \theta),$$
$$e^{-bx} = \frac{(a - (1 - \theta))}{d},$$
$$\ln(e^{-bx}) = \ln((a - (1 - \theta))/(a - 1)),$$
$$b = \frac{1}{x} \ln\left(\frac{a - (1 - \theta)}{a - 1}\right).$$

This demonstration is also valid for the production function associated with MTCT.

References