UNAIDS Epidemiology Reference Group

Recommended methodology for the estimation and projection of HIV prevalence and AIDS mortality in the short-term

Based on a meeting held at La Mainaz Hotel, Gex, France
January 2001

Table of Contents

Introduction

This meeting was held to develop recommendations for the methodology that will be used by UNAIDS to produces the May 2001 estimates and projections of HIV prevalence and AIDS mortality for all countries. The list of participants is given at the end of this document. The meeting was planned at the Epidemiology Reference Group meeting, held in Frascati in October 2000. At that meeting the following recommendations were made for the short-term projection of HIV prevalence:

- Keep using available ANC data – time series of prevalence data given as a proportion of those aged 15-49.
- Don't include age-structure in the model used for short-term projections
- UNPOP to translate non-age-structured projections into a pattern of prevalence and incidence with respect to age for projecting demographic impact
- Move on from the `gamma`- Epimodel - try different epidemiological functions (or set of equations) where the parameters of model have a biological/epidemiological/demographic meaning
- Different models should be validated using time series data of ANC data in four ways:
    i. Goodness-of-fit
    ii. Predict 'future' – edit out last 5-years of ANC data, use the model to project HIV prevalence over that period and then compare
    iii. Cross-validation of projections made by different models
    iv. Scant and poor input ANC data should be reflected in large confidence intervals about projected prevalence (or a refusal to project based on poor data). This is aimed at preventing model misuse.
- Members of reference group to do the above validation work and to meet and decide on the function/set of equations to be used for the short-term projections at a meeting in Geneva just before or after Christmas.
- The interface developed by Tim Brown will be used to implement the chosen function/set of equations. This model adds power to the method by allowing prevalence fits for different cohorts [sic. population subgroups] within a country and the possibility of fitting ranges rather than point estimates of prevalence
- The possibility of fitting ranges of data, and also dividing data into cohorts [sic. population subgroups], should be given some more consideration.

The full set of recommendations from the meeting in Frascati is available at http:// www.ceid.ox.ac.uk/unaids.

This meeting therefore followed up on these recommendations through the hard work of a group of epidemiologists, demographers and statisticians. Each participant presented the approach to short-term projection that they had been working on, after which discussion led to recommendations for the model to be used by UNAIDS in the 2001 estimates and projections. Following the meeting further work was carried out to validate the proposed model. This report presents a summary of the presentations given at the meeting, the issues that were discussed and a proposal for the model to be used for the 2001 UNAIDS country-specific estimates and projections of HIV prevalence. A formal description of this model is given at the end of the report and some analyses of its behaviour presented.

The models presented at the meeting

Six different approaches to the estimation and projection of HIV prevalence were presented at the meeting. Although somewhat divergent, these approaches had epidemiologic and demographic realism as their underlying theme, and reassuringly gave very similar prevalence projections despite their diversity. With the exception of one approach, projections were for the adult population without stratification by age. An additional model described in a manuscript distributed by Griff Feeney is available at http:// www.ceid.ox.ac.uk/unaids, together with a note on the sampling error of national prevalence estimates derived from HIV sentinel surveillance sites.

A brief summary of the models presented at the meeting follows. In order to simplify the summary, a standard notation is introduced: Z is the number of individuals at-risk, X is the number of individuals not at-risk, and Y is the number of HIV positive individuals. The total population size, $N = X + Y + Z$. The force of infection, $\lambda = rY / N$, where r is a constant defining the relative magnitude of λ.

John Stover of the Futures Group International, Glastonbury, CT, USA presented a simple epidemiological model with demographically realistic rates of entry into and exit from the adult population. This model stratifies the population into 'susceptible', 'infected' and 'AIDS', with movement between these categories defined by a set of ordinary differential equations (ODEs). Progression rates are assumed constant and correspond to the reciprocal of the mean duration of infection prior to AIDS and the mean time spent with AIDS. To capture the rapid declines in HIV prevalence seen in some African countries, the force of infection λ, is not only dependent on HIV prevalence but also the cumulative number of deaths due to AIDS. Thus,

$$\lambda = r(Y / N).(D / N).M$$

where D is the cumulative number of deaths due to AIDS and M is a scaling factor for behaviour change.
Entry into the adult population can be calculated as the crude adult birth rate 15 yrs ago multiplied by the probability of survival to age 15. This is complicated by fertility reductions due to HIV and vertical transmission of HIV. This rate of entry can be approximated by the number of 15 yr olds divided by the population age 15-49 yrs. The non-AIDS adult death rate is determined from model lifetables, although this will be an approximation due to changes in the age structure of the adult population caused by AIDS mortality.
The fits to median urban and rural HIV prevalence data for Kenya and Malawi using this simple epidemiological model were found to agree with earlier projections made by the Futures Group, based on detailed analyses of the data and more complex models.

Geoff Garnett and **Nick Grassly** from the UNAIDS Epidemiology Reference Group secretariat at Imperial College, London University presented a simple epidemiological model, with an explicit demographic parameterisation of the rate of entry into and exit from the adult population (these demographic rates were taken directly from Basia Zaba's work). This model divides the population into three categories: not at-risk, at-risk and infected, with movement between these categories defined by a set of ODEs. Progression from infection to death was assumed to occur at a constant rate corresponding to the reciprocal of the mean duration of infection. The force of infection

is dependent on the number of HIV positive individuals. The introduction of a not at-risk population meant that observed patterns of HIV epidemics, where the epidemic can peak at a low prevalence, could be captured without the need to invoke behaviour change.
As the HIV epidemic progresses, AIDS mortality results in a decline in the at-risk population relative to the not at-risk population. In other words the prevalence of risky sexual behaviour declines. An additional parameter in this model ϕ was introduced to allow an increase in the rate of entry into the at-risk population, relative to the not at-risk population, in response to this decline. A large ϕ corresponds to maintenance of levels of risky behaviour in the face of AIDS mortality, whilst $\phi = 0$ implies there is no change in the relative rate of entry to the at-risk population during the course of the epidemic. This parameter directly determines the endemic prevalence of HIV following the initial peak.
The fitting of individual sentinel surveillance site data for a given country using maximum likelihood was demonstrated to give more stable fits than attempts to fit median prevalence estimates using least squares. The use of likelihood ratio theory to obtain statistical confidence limits about estimates and projections of HIV prevalence was described. These confidence limits are responsive to changes in the quantity and quality (sample size, but not representativeness) of the sentinel surveillance site data.

Ping Yan from the Bureau of HIV/AIDS, STD and TB at Health Canada utilised an explicit functional form of an epidemic, based on a convolution of a log-logistic curve for new infections and a survival function post-infection such as the Weibull, log-logistic or Gamma. The log-logistic function for new infections was parameterised in such a way as to allow only a fraction of the population to be at-risk, and has been previously demonstrated to provide a close approximation to simulated stochastic epidemics.
Using this approach it was possible to fit HIV prevalence data to demonstrate the sensitivity of estimates of HIV incidence and prevalence to assumptions about the survival function. In general it was noted that prevalence projections were not very sensitive to the assumed survival function, whether log-logistic, gamma or Weibull, but that the corresponding incidence needed to generate these prevalence curves could be very different. Thus estimates of incidence should be interpreted with caution.

Marc Artzrouni from the Laboratory of Applied Mathematics at the Université de Pau et des Pays de l'Adour, France presented an epidemiological model applied to a population assumed to be growing exponentially. Individuals were defined as susceptible, infected and non-susceptible, again allowing peak prevalence values at reasonable levels even for rapidly spreading epidemics. Progression from infection to death was modelled using a Weibull function.

David Schneider from Life Assurance, Botswana presented a simplified version of the Actuarial Society of South Africa's 'ASSA 2000' model of HIV spread. This model stratifies the population by age, sex and risk group, where risk groups correspond to prostitutes, STD patients, at-risk and not at-risk individuals. This model was able to fit observed prevalence data, but the large number of parameters meant there was not a unique set of parameters that described the best fit. For this reason, and because of time constraints envisaged in the country-specific curve fitting procedure, prevalence data was fitted by eye.

This model is used by ASSA to model risk of death by age, sex and risk group. Such information is then used in the design of insurance policies.

Basia Zaba from the Centre for Population Studies at the London School of Hygiene and Tropical Medicine presented an epidemiological model with rates of entry into and exit from the adult population based on demographic calculations. The adult population is divided into susceptible and infected individuals, with the force of infection defined as a function of HIV prevalence,

$$\lambda = rY / N$$

where r can change over time as a piecewise linear function. By allowing r to decline after a given time, the observed patterns of HIV epidemics can be fitted, with rapid initial spread but peak prevalence at low values.
Different methods for the calculation of the rate of entry to the adult population were described, and the use of the crude adult birth rate with the probability of survival to age 15 found to be the most consistent estimator of the rate of entry (cf. the 15^{th} birthday rate).
Different expressions for the rate of progression from infection to death were explored. By allowing the rate of progression to depend on the mean time since infection in the population, the relationship between the probability of dying and time since infection was more accurately reflected than by assuming a constant rate of progression. Defining progression rates in this way also obviates the necessity of using integro-differential equations if time since infection is to be explicitly represented. The projections of HIV prevalence that assumed a rate of progression dependent on mean time since infection tended to better fit the observed data than projections that assumed a constant rate of progression.
The instability of fits to median prevalence data only, rather than individual site data, was demonstrated.

The key features of the approaches to estimating and projecting HIV prevalence described above are captured in Table 1. This table provides a more detailed summary of the models presented.

Table 1: A summary of the features of the models presented at the meeting. Models are referred to by the name of the presenter.

	John Stover	Geoff Garnett & Nick Grassly	Ping Yan	Marc Artzrouni	David Schneider	Basia Zaba
1. ODEs or parametric curve	ODEs	ODEs	Parametric	ODEs	ODEs	ODEs
2. End date of projections	2000	2010	2005	2010/20	2050	2030
3. Age- and/or sex-stratified	No	No	No	No	Yes	No
4. HIV positive survival	Exponential	Exponential	Gamma, Weibull or log-logistic survival	Weibull	Weibull	AIDS mortality a function of mean duration of infection in the population
5. 'Not at-risk' population included?	No	Yes	Yes (scale parameter, c)	Yes	Yes – 'at-risk', 'not at-risk', prost., STD	No
6. Requires population level behaviour change to fit observed prevalence trends	Yes – the force of infection decreases with increased AIDS mortality	No (although behaviour change can be specified by setting phi >0)	No	No	No	Yes – force of infection a piecewise linear function of time
7. Demography captured	Yes	Yes	N/A	Exponential population growth at fixed rate	Yes	Yes
8. Rate of exit of adult population (implications for adult population modelled i.e. 15+ vs. 15-49)	Crude adult death rate for 15-49	Crude adult death rate for 15+	N/A	N/A	Age-specific non-AIDS death rates allows 15+ or 15-49 population to be modelled	Crude adult death rate for 15+
9. Rate of entry to adult population	15^{th} birthday rate or births and survival to age 15	Births and survival to age 15	N/A	N/A	Explicit age-structure	Births and survival to age 15

Table 1 continued…

	John Stover	Geoff Garnett & Nick Grassly	Ping Yan	Marc Artzrouni	David Schneider	Basia Zaba
10. Data fitted	1) Sites with ≥ 10 yrs of data 2) median urban and rural fitted separately - national estimate is a weighted average	1) National level data as a weighted average of urban/rural median prevalence 2) All sentinel sites accounting for sample size	Average prevalence for groups of sites classified by geography	Rural and urban median prevalence	National level estimates as a weighted average of urban and rural	National level estimates as a weighted average of urban and rural
11. Curve fitting procedure	Least squares	1) Least squares 2) Maximum likelihood	Least squares	1) Least squares 2) eyeball	Eyeball	Least squares
12. Stability of fits examined or statistical confidence limits	No	Yes	Yes	No	No	Yes

Recommended features of the UNAIDS model for the 2001 country-specific estimates and projections of HIV prevalence and incidence

1. ODEs or parametric

The model used will be based on differential equations that describe the movement over time of people between different categories (e.g. susceptible, infected, etc…).

2. Timescale

The timescale for the projections will be flexible, with the model capable of producing projections for between 5 and 50 years.

3. Age- and/or sex-stratified

As decided previously based on data constraints and the need for a practical model, there will be no stratification of the model into age and sex categories. Ongoing work will explore sex and age stratification and how to link the short-term model to a fully demographic model. For the 2001 estimates of age-specific HIV incidence needed for demographic calculations, the age and sex-stratified model *'Spectrum'*, developed by the Futures Group International, will be used.

4. HIV positive survival

The Weibull distribution will be used to reflect the pattern of AIDS mortality in relation to time since infection. The user of the model will be given the choice of one of three median survival times, which correspond to three predefined categories of progression (fast, medium and slow). A formal review of the literature will be carried out and discussion groups used to define these rates of progression.

5. 'Not at-risk' population included?

Yes, a not at-risk population will be included.

6. Requires population level behaviour change to fit observed prevalence trends

The incorporation of a not at-risk population in the model allows the model to capture the dynamics of observed HIV epidemics without the necessity to invoke sexual behaviour change. Specifically, the model can reflect an epidemic that peaks at a low prevalence, but where the spread of HIV from a prevalence of zero to the peak occurred over just a few years. Against this picture of the 'natural course' of the epidemic, model parameters can be changed to reflect either a spontaneous change in sexual behaviour in response to observed AIDS mortality, or decreased rates of transmission and reduced levels of risky sex due to effective interventions.

7. Demography captured

It was agreed that the rates of entry into and exit from the adult population due to births and non-AIDS related deaths respectively should be defined using available demographic information. The data required for these demographic calculations will be supplied by the UN Population Division (Hania Zlotnik). The following two recommendations specify the details of these demographic calculations.

8. Rate of exit of adult population

Prevalence data from antenatal clinics (ANCs) acting as sentinel surveillance sites is reported for the age-range 15-49 yrs. If the model to be implemented by

UNAIDS is to contain demographic realism, then the definition of the rate of entry and exit from the adult population has implications for the age range of the population modelled. The rate of entry can be defined as the 15th birthday rate, derived from the crude adult birth rate 15 yrs earlier, and the probability of survival from birth to age 15. Thus the lower age boundary of the population modelled is 15. If the upper age of the population is to be 49, then the rate of exit from the adult population (excluding AIDS mortality of HIV positives) needs to include both non-AIDS mortality rates for the 15-49 population and the rate of aging out of the population (50th birthday rate). Because AIDS mortality can result in substantial changes in the age structure of the adult population, which is not captured in the non-age-stratified model described here, defining these rates will be complicated.

At the meeting two approaches to the problem were suggested. The first was based on modelling the open-ended 15 yrs upwards (15+) population, such that only the crude adult death rate would need to be specified, and adjusting prevalence data to reflect this age distribution. It was hypothesised that the level and trends in this crude adult death rate in response to a shifting in the adult age structure due to AIDS mortality would be easier to capture than both the death rate and the 50th birthday rate. The second proposal was to estimate the 50th birthday rate by using an approximation based on the mean age at infection in the adult population.

Before considering these two approaches it should be noted that if we are interested only in percentage HIV prevalence and incidence, then their fit to data from sentinel surveillance sites is robust to changes in how the rate of exit is specified (Figure 1). It is clear that when fitting prevalence data from all sentinel sites, the fitted prevalence and incidence are very similar for all values of μ.

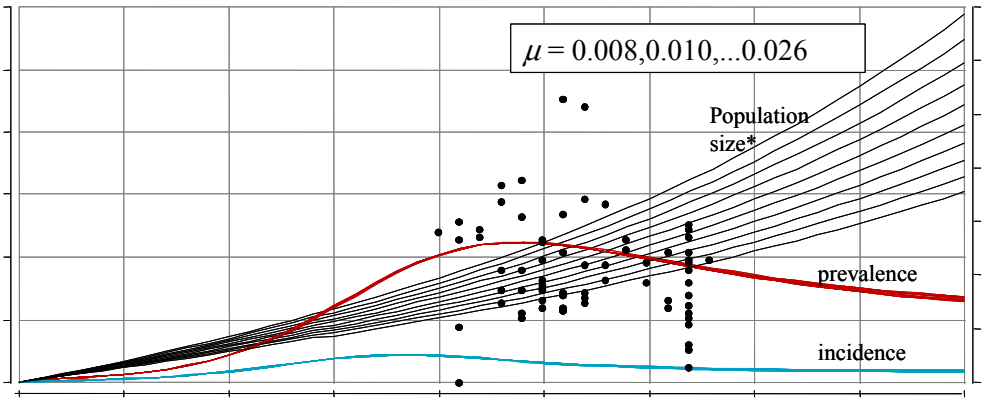


Figure 1 Maximum likelihood estimates of HIV incidence and prevalence, and the corresponding population projections for DR Congo sentinel surveillance site data for 10 different values of the rate of exit from the adult population μ. The range of μ used is for that observed in sub-Saharan Africa.

Therefore it may not be too important how μ is defined, and we can continue with the definition of the adult population as being 15-49 yrs. However, as Figure 1 makes clear, if we are interested in projecting population numbers, or *numbers* of HIV positives,

incident cases or AIDS deaths, then μ needs to be specified in a demographically correct way.

Two approaches to the definition of the rate of exit:
1). *Crude adult death rate*. This is the crude rate of mortality for the 15+ population in the absence of AIDS, which can be estimated from country-specific life tables in the years prior to the AIDS epidemic. The procedure for calculating the crude adult death rate was described by Basia Zaba. If N(x, t) is the size of the population at time t, for age group x, then using the UN Population division estimates of population numbers by 5 year age group, we can calculate the crude adult death rate over time t to $t + 5$, as

$$\mu = 0.5\left[-0.2\ln\left(\frac{N(15+,t+5)}{N(10+,t)}\right)-0.2\ln\left(\frac{N(20+,t+5)}{N(15+,t)}\right)\right]$$

This calculation might be carried out for the years 1950 to 1980 and projected using a linear model, or just the most recent estimate prior to the HIV epidemic used (e.g. 1980). The crude adult death rate will be an approximation since it will change through time due to changes in the age structure of the adult population caused by AIDS mortality. Further investigation into the nature of these changes would be useful.

Using the crude adult death rate in this fashion implies that the adult population which is modelled includes all individuals 15 yrs old and over. Because sentinel surveillance data is for pregnant women in the age range 15-49 yrs, it may be necessary to adjust this data prior to fitting the model. Further work following the meeting at La Mainaz suggests however, that such adjustment may be unnecessary. For most countries with large HIV epidemics the population is growing and life expectancy is low. This means that most people are under 50, and HIV prevalence for the open-ended 15+ age group will not be very different from the 15-49 age-group. Some simulations using an age- and sex-stratified model of HIV prevalence confirm this when applied to Togo and Uganda (Figure 2). This model stratifies the force of infection from the UNAIDS model by age, using the age-profile of the force of infection from cohort studies in Masaka, Kagera, Kisesa, and Mwanza (data courtesy of Basia Zaba).

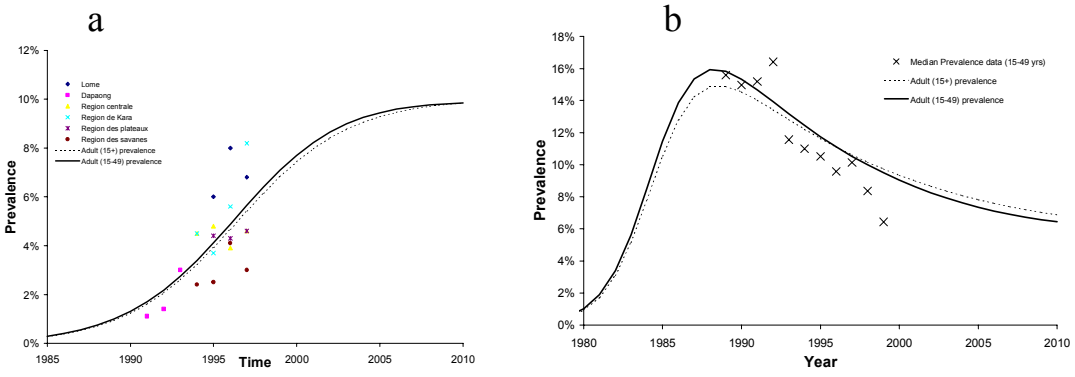


Figure 2 Projected prevalence from an age and sex-stratified model for adults of age 15-49 and for those 15+ for fits to data from a) Togo and b) Uganda.

It is usually assumed that prevalence for the 15-49 age range is higher than for the 15+ population, because death from AIDS usually occurs before age 50 and HIV incidence for the 50+ population is low. This is seen for the prevalence projections for Togo, although the difference can be seen to be very small. The same is seen for Uganda until 1997 when the pattern is reversed. These results are clarified in Figure 3, which shows the absolute and percentage difference in prevalence in the 15+ and 15-49 age-groups over the course of the epidemic in Togo and Uganda. The difference between the prevalence for the two age groups is small and never exceeds 10% of the mean or 1-2 prevalence percentage points.

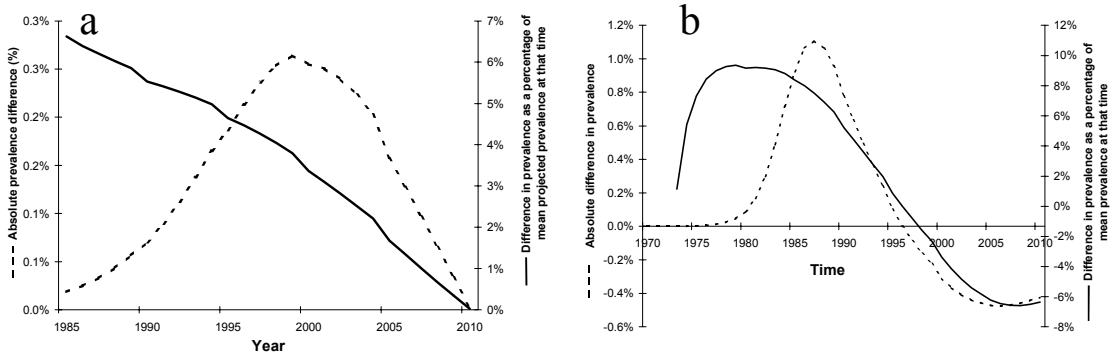


Figure 3 The absolute (dotted line) and percentage (solid line) difference in prevalence between the 15-49 and the 15+ age groups for age and sex-stratified projections fit to data from a) Togo and b) Uganda.

The fact that the prevalence for the 15-49 yrs age group may drop below that for those 15+, as for the Ugandan projections, is a result of a rapid decline in incident HIV cases (concentrated amongst the younger age-groups), and aging of those people with HIV. The same pattern is observed, even if AIDS free survival times are shorter for older individuals (results not shown).

2). *Crude adult death rate for the 15-49 yr age-group plus rate of exit from this age-group* This would enable the model to explicitly reflect the 15-49 age group, but is complicated by changes in the age structure of the adult population through time caused by AIDS mortality. The impact of such changes on a rate of exit, μ, that includes the 50^{th} birthday rate are likely to be more dramatic than when μ is simply taken as the crude adult (15+) death rate. However this impact may be captured in a model relating μ to the mean age at infection in the adult population. The possibility of describing such a model is to be further investigated by John Stover.

In the short-term, given time constraints, it was decided to proceed with the former approach of using just the crude adult death rate for the open-age group to define μ when modelling heterosexual epidemics. Further consideration needs to be given to epidemics amongst other risk groups, such as intravenous drug users (IDUs). The simple epidemiological model, as recommended here, should be an adequate reflections of the spread of HIV amongst IDUs by needle sharing, but the demographic parameters are clearly different for these groups. The rate of exit will need to be defined with reference both to mortality amongst IDUs, which tends to be higher than in the general

population, and rates of giving up intravenous drug use. This may be derived form the mean time spent injecting drugs, if known.

9. Rate of entry to adult population

For heterosexual epidemics the adult population size will be explicitly modelled, with the starting value being based on UN Population Division estimates for the starting year of the model projections. Using the crude adult fertility rate based on UN Population Division data and survival of HIV negative children to age 15 based on the country lifetables, the number of individuals entering the adult population each year can be specified. This procedure is described in detail in the formal model description at the end of this document. It is assumed that all children who acquire HIV via vertical transmission do not to survive to age 15.

For IDUs the rate of entry may be derived from the rate of exit as defined previously, and knowledge about trends in the size of the drug injecting population. Given the uncertainty about these parameters, estimates of the number of HIV positive IDUs should be given with the necessary caveats. Percentage prevalence and incidence estimates are likely to be more accurate.

10. Data fitted

All data from individual sentinel surveillance sites will be used rather than median prevalence estimates, thus avoiding to a certain extent the problem of unstable fits due to the small number of fitted data points (Figure 4). Different prevalence curves can be fitted to different user-defined groups of sites e.g. urban, rural, and coastal.

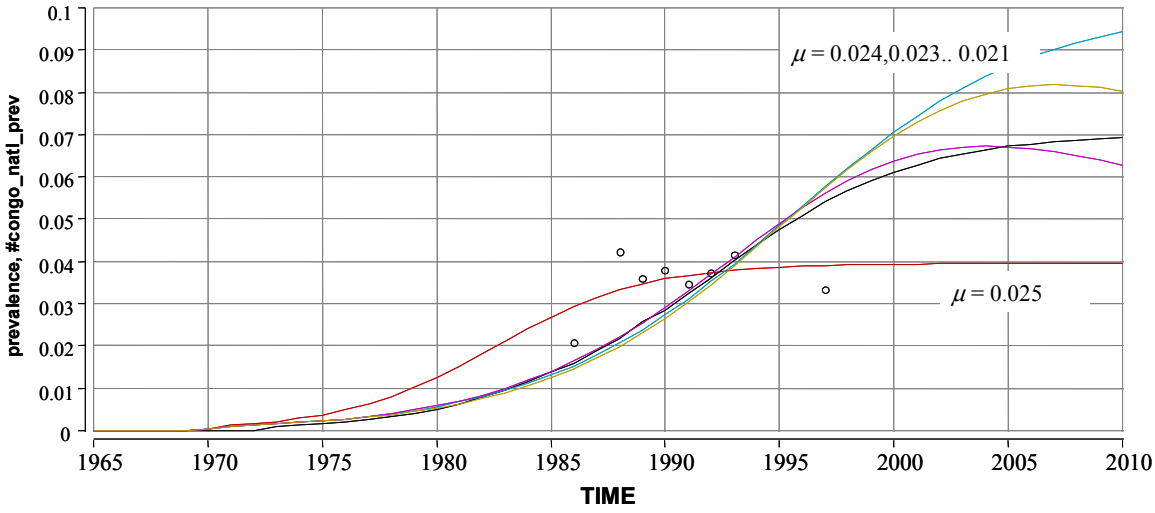


Figure 4: Different prevalence projections obtained for the Democratic Republic of Congo for slightly different demographic parameters when only median prevalence data is fitted. Fits based on all sentinel surveillance site data are more stable to changes in demographic parameters (Figure 1).

11. Curve fitting procedure

It was decided to define the groups of sites to be fitted (e.g. rural, urban, national) prior to the curve fitting procedure. The prevalence curve for that group can

then be fitted to all *individual sites* using the least squares approach. Care should be taken that local optima are not found by the fitting procedure.

This least squares fitting procedure should be weighted, such that sites considered outliers or 'unbelievable' can be weighted down. In addition, sites that are thought to be particularly representative can be weighted up. In countries where enough information exists, population proportional sampling (PPS) may be considered.

12. Statistical confidence limits

These will not be produced for the May 2001 estimates and projections. However, it was decided that they will be necessary in the next year or two and therefore need to be explored more fully. It is important that the relative magnitude of the statistical component of error, which is likely to be small, compared to the impact of sex differences in prevalence, selection biases in the sample, and mis-specification of model is made clear. It is also important to differentiate between the confidence about estimates of prevalence based on existing data, and confidence in projections of prevalence into the future. A note on sampling error of estimates of HIV prevalence from sentinel surveillance sites prepared by Griff Feeney is attached as an appendix.

13. Discrete vs. continuous time implementation

The model will be implemented in continuous time (i.e. solved in very small time steps on a computer).

14. Sex ratio of prevalence projections

This will be specified after fitting the model and be based on empirical estimates. It will be allowed to vary over time if enough information is available.

15. Output required from the model

The outputs required from the model are:

1. Sex-specific prevalence for age groups 15-49, and 0-14. The latter estimate will be produced by projecting the number of HIV positive and negative births, and survival curves for HIV positive and negative children (the survival of HIV negative children will be country-specific and based on UN Population Division data).
2. Incidence for the same age-groups also by sex, where incidence is described as the number of new infections per year per individual in the population. It may also be useful for an incidence rate per susceptible per year to be output (traditionally termed the force of infection).
3. Numbers of adult and child AIDS deaths by sex

Formal Model Description

Z = at-risk population
X = not at-risk population
Y = infected
$N = X + Y + Z$

$$\frac{dZ}{dt} = f(X/N).E_t - (\mu + rY/N + \iota)Z \qquad \ldots(1)$$

$$\frac{dX}{dt} = (1 - f(X/N)).E_t - \mu X \qquad \ldots(2)$$

$$\frac{dY}{dt} = (rY/N + \iota)Z - \int_0^t (rY_x/N_x + \iota_x).Z_x.g(t-x)\mathrm{dx} \qquad \ldots(3)$$

where $f(X/N)$ is the fraction of those individuals entering the adult population (E_t) who enter the at-risk group Z, and is given by

$$f(X/N) = \frac{\exp\left[\phi(\frac{X}{N} - (1 - f_0))\right]}{\exp\left[\phi(\frac{X}{N} - (1 - f_0))\right] + \frac{1}{f_0} - 1} \qquad \ldots(4)$$

where f_0 is the fraction of individuals entering the at-risk group at the start of the HIV epidemic, and ϕ is a parameter which determines how $f(X/N)$ changes with respect to f_0 once AIDS mortality starts to occur. For $\phi = 0$, $f(X/N)$ remains constant at f_0. For $\phi > 0$, the at-risk population is maintained in the face of AIDS mortality by an increase in the proportion of people entering the at-risk group. This might be considered a reflection of the demand for risky sex maintaining the size of the at-risk population, or the gradual exposure of previously isolated (and therefore not susceptible) risk groups to sources of HIV infection. This isolation may be geographic or cultural. For $\phi < 0$, AIDS mortality in the at-risk group results in a decrease in the proportion of individuals entering this group. This negative feedback means that observation of AIDS mortality makes people less likely to engage in risky sexual behaviour. The relationship of $f(X/N)$ to X/N (the fraction currently not at-risk in the population) and ϕ is shown in Figure 5a. The impact different values of ϕ can have on prevalence projections is shown in Figure 5b. These projections show that ϕ is a determinant of the endemic prevalence of HIV.

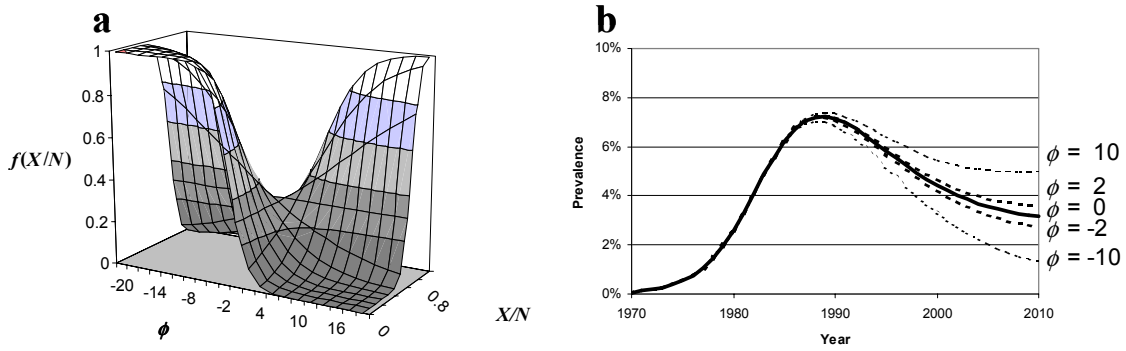


Figure 5 a) The relationship of $f(X/N)$ to X/N and ϕ. For positive ϕ, if the at-risk population declines (and hence fraction not at-risk X/N increases), the fraction recruited to the at-risk population increases, asymptotically approaching 1. For negative ϕ the opposite patterns is observed, whilst for $\phi = 0$, $f(X/N)$ remains constant at f_0 (which for this graph is set to 0.3). b) The impact of different values of ϕ on prevalence projections.

$g(x)$ is the density function for mortality from AIDS or other causes, and is given by,

$$g(x) = (\mu + \alpha x^{\alpha-1} / \beta) \exp[-\mu x - (x/\beta)^{\alpha}] \qquad \ldots(5)$$

where α is the shape parameter of the Weibull distribution fitted to HIV survival times and β is the position parameter. The parameter β can be defined with reference to the median survival time, m:

$$\beta = m / [\ln(2)^{1/\alpha}] \qquad \ldots(6)$$

For ease of use of the model it may be best to pre-define the shape parameter α based on available empirical data, and use three values of median survival times corresponding to slow, medium and rapid progression.

Model parameters

Calculated or fixed model parameters

1. crude adult (15+) death rate, μ
2. numbers entering adult population at time t, E_t (dependent on births and survival to age 15 - see below)
3. force of mortality due to AIDS, x years after infection (Weibull function for survival used)

When modelling a heterosexual epidemic the numbers entering the adult population at time t, E_t can be specified in terms of the births of HIV negative children B^{-}_{t-15} occurring 15 years previously, and the cohort survival proportion, l to age 15. In turn, B^{-}_{t-15} can be expressed in terms of the birth rate b that is applied to the adult population

at the time, with suitable allowances made for the probability of vertical transmission, υ, and a fertility reduction term ε applied to the HIV positive population:

$$E_t = B^-_{t-15} \cdot l \qquad \cdots(7)$$

$$B^-_{t-15} = b \cdot \left[X_{t-15} + Z_{t-15} + (1-\nu) \cdot \varepsilon \cdot Y_{t-15}\right] \qquad \cdots(8)$$

We make the assumption that HIV positive births, B^+, do not survive to adulthood. However, if we are interested to project their number, this is simply:

$$B^+_t = \nu \cdot \varepsilon \cdot Y_t \qquad \cdots(9)$$

To ensure compatibility between the initial birth rate and the fifteenth birthday rate at the start of the projection, we assume an initially stable age structure and equate the initial growth rates, c, in the adult population and the general population. This involves solving the following, using the base year (e.g. 1965) values of b_{1965}, l_{1965} and μ_{1965}.

$$b \cdot l_{1965} \cdot e^{-15c} - \mu_{1965} - c = 0 \qquad \cdots(10)$$

Model parameters estimated when fitting prevalence data

1. Fraction of population entering at-risk group before epidemic starts, f_0
2. Summary measure of sexual contact rates and transmission probabilities, r
3. Start date of the epidemic - specified by having the 'exogenous' force of infection, ι, a function of time
4. Response of at-risk population recruitment to AIDS mortality, ϕ

Model behaviour

The model produces a prevalence curve rising initially exponentially and then asymptotically approaching an epidemic peak, before potentially declining to an endemic prevalence. Because of the specification of the Weibull survival function of HIV positives, the decline to an endemic prevalence follows a pattern of damped oscillations. A fit to data from the Democratic Republic of Congo and projected prevalence until 2050 demonstrates these oscillations (Figure 6).

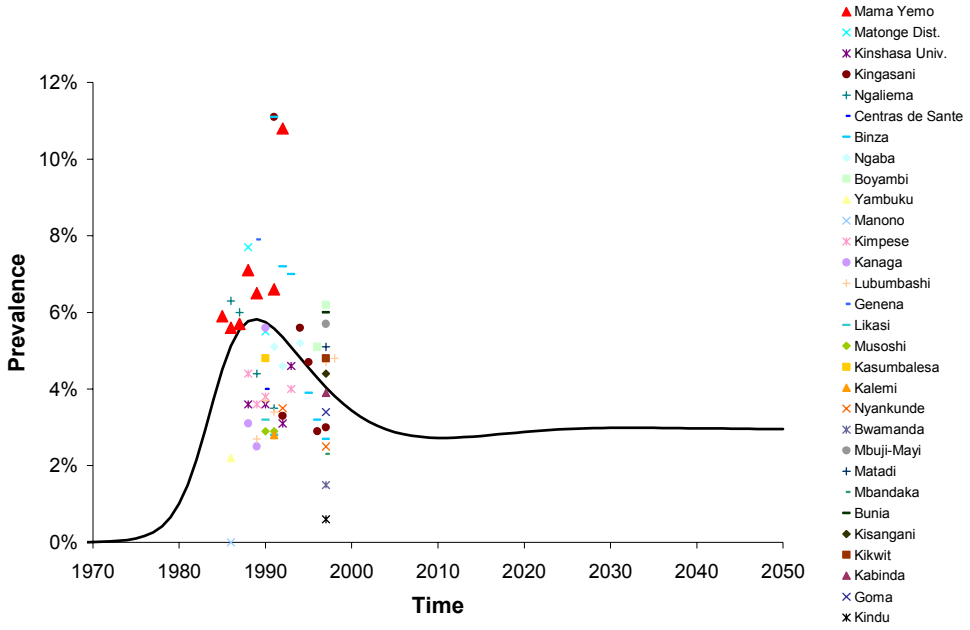


Figure 6 Fit of UNAIDS 2001 model of HIV prevalence to DR Congo sentinel surveillance site ANC data

The amplitude of these oscillations is determined by the sharpness of the incidence peak, which in turn is determined by r and f_0. For most fits to country sentinel surveillance site data these oscillations were of a small amplitude.

The sensitivity of the prevalence during the course of the epidemic to a small increase in the epidemiological parameters of the model is shown in Figure 7. The constant multiplier in the force of infection, r, and the initial proportion of the population at-risk, f_0, determine the rate of take-off of the epidemic. The size of the peak of the epidemic is insensitive to changes in r, but sensitive to the fraction of the population at-risk f_0. The endemic prevalence is determined by all three epidemiological parameters, but with sensitivity to ϕ restricted to this part of the prevalence curve, such that ϕ can be used to determine the final endemic prevalence.

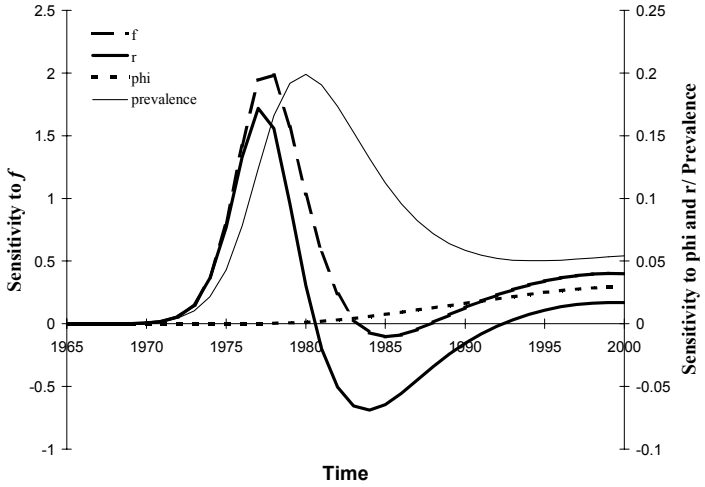


Figure 7 Sensitivity of prevalence through the course of the epidemic shown to small increase in the epidemiological parameters r, f and ϕ. Sensitivity is defined as S = [P(x+dx) - P(x)] / dx, where P(x) is the prevalence for parameter with value x, and dx is the small change in x.

When fitting sentinel surveillance site data, the projected prevalence is sensitive to the specification of the force of mortality from AIDS following infection with HIV (in contrast to the insensitivity of fitted prevalence curves to the specification of the demographic parameters μ and the birth rate). The fit of DR Congo sentinel surveillance site prevalence data where an exponential and Weibull distributed survival with HIV are hypothesized is shown in Figure 8a (in both cases median survival was 10 yrs). Although the difference between the two prevalence projections is not great, it becomes more pronounced for projected prevalence further into the future.

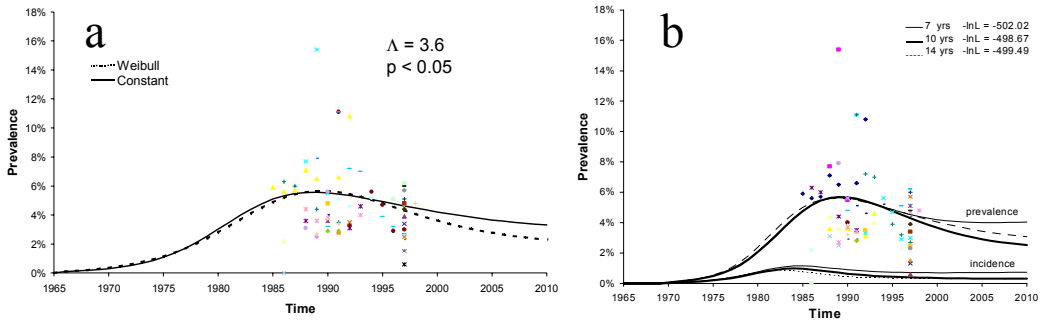


Figure 8 The maximum likelihood fit of prevalence curves to sentinel surveillance site data from the DR Congo for **a)** an exponential and Weibull distributed HIV positive survival time, and for **b)** Weibull distributed survival with median survival times of 7, 10 and 14 yrs.

Figure 8b shows the fit to DR Congo prevalence data for different assumptions about the median survival time post-seroconversion for HIV positives. Again, somewhat different fits are obtained for different survival times[1]. Thus, correct specification of the survival function and median survival time is likely to be important if better prevalence estimates and projections are to be produced.

In order to estimate ϕ and hence the endemic prevalence of HIV after the epidemic has peaked, some prevalence data post this peak are required. For countries where such data exists, ϕ may be reasonable well specified. For example, the likelihood surface around the maximum likelihood estimate (MLE) of ϕ for the Democratic Republic of Congo is steep enough to obtain upper and lower bounds on ϕ using the likelihood ratio statistic (Figure 9a). In contrast, for Togo, where the prevalence of HIV is still rising, ϕ can take almost any value without any significant impact on the likelihood (Figure 9b). For countries where the epidemic has yet to peak it may be advisable to choose a value of ϕ *a priori* – whether this should be 0 or some other value remains unclear.

[1] Interestingly, the best fit of these three mean survival times is 10 years, consistent with our expectations. However, cohort studies provide better information on median survival than do maximum likelihood estimates from prevalence data.

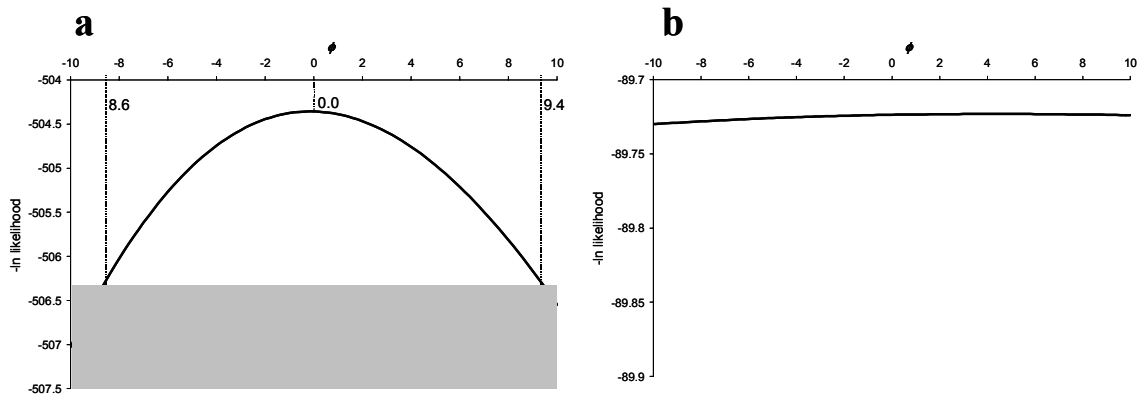


Figure 9 Likelihood surfaces about the MLE of ϕ for a) the Democratic Republic of the Congo and b) Togo. For the DR Congo upper (9.4.) and lower (-8.4) confidence limits about ϕ based on the likelihood ratio statistic are shown.

List of participants

Name		Affiliation
Marc	Artzrouni	Lab. Math. Appl., Université de Pau et des Pays de l'Adour, France
Tim	Brown	Population and Health Studies, East-West Center, Honolulu
Griff	Feeney	Demographer, Honolulu
Geoff	Garnett	UNAIDS Epidemiology Reference Group Secretariat, London
Peter	Ghys	UNAIDS
Nick	Grassly	UNAIDS Epidemiology Reference Group Secretariat, London
Stefano	Lazzari	UNAIDS
David	Schneider	Actuarial Solutions, Botswana
Karen	Stanecki	US Bureau of the Census
John	Stover	Futures Group International, CT, USA
Neff	Walker	UNAIDS
Peter	Way	US Bureau of the Census
Ping	Yan	Health Canada
Basia	Zaba	London School of Hygiene & Tropical Medicine, UK
Hania	Zlotnik	UN Population Division

APPENDIX -Note on the Sampling Error of National Prevalence Estimates Derived from HIV Sentinel Surveillance Data

15 January 2001
Griffith Feeney gfeeney@gfeeney.com

Prepared for the La Mainaz Meeting on Short Term Projections,
UNAIDS Epidemiology Reference Group
25-26 January 2001, La Mainaz Hotel, Mijoux, France

Abstract *HIV Sentinel Surveillance (HSS) data may be regarded as an imperfect realization of a two stage sample for which sampling error can be calculated. The calculated sampling errors are evidently lower bounds for the actual sampling error of national HIV prevalence estimates derived from HSS data. For a simulated population with 32.5 percent prevalence, samples of 300 women from each of 10 antenatal clinic (ANC) sites give a sampling distribution with an inter-quartile range of 2.6 percent. Estimated prevalence would therefore be expected to err by more than 1.3 percent half the time. It is suggested that, where numbers of HSS sites vary greatly from one year to another, fits of projection models should take account of the variable sampling error of the data points for different years.*

§1 Introduction HIV Sentinel Surveillance (HSS) was not intended to be a basis for national estimates of HIV prevalence. For most countries, the antenatal clinic (ANC) sites are not a representative sample of any well-defined population. Self selection of women for clinic attendance makes representative sampling of women impossible. Sampling error therefore cannot be established by standard methods.

At the same time, it is clear that (i) national prevalence estimates derived from this data are subject to sampling error and (ii) that this error is higher than it would be if sites and women *were* representative but otherwise comparable samples.

The idea of this note is to (a) specify a two stage, representative sample that would provide data similar to that provided by HHS data; (b) calculate sampling errors for national prevalence estimates based on this sample; and (c) regard these errors as lower bounds for the sampling error of national estimates derived from HHS data.

While it may be possible to compute sampling distributions for some estimators directly [1], the method used here is to simulate "true" data for an hypothetical country, draw repeated random samples from this data, and compute estimates from these samples to obtain sampling distributions.

Estimates of sampling error are pertinent to the fitting of models to observed prevalence because, in general, sensible fits require some notion of possible error in the data fitted to. Perhaps more importantly, the number of HSS sites available in any particular

country varies greatly from year to year, and this implies that data points for different years should be accorded different weights when fitting models.
If least squares fits are used, for example, *weighted* least squares fits with weights varying inversely with estimated sampling error, will evidently give far more defensible results than unweighted least squares.

§2 Designing the Sample Imagine a particular country divided into n well-defined geographic areas, each corresponding to the catchment area of an antenatal clinic (ANC). Suppose that every such area does in fact contain an ANC. Suppose further that a simple random sample of k of these n sites is drawn, and that from each of these k sites a simple random sample of m_k pregnant women is drawn.

§3 Selecting an Estimator National prevalence estimates seem to be mostly based on medians of site-specific prevalence values. Why is the median used in preference to the obvious alternative, the "binomial" estimator defined as the sum of the HIV positive women from all sites divided by the sum of the number of women tested at each site? Is it because of the robustness of the median against outliers? Because numbers of women tested are sometimes not available? The question is pertinent because the binomial estimator is arguably superior. Both estimators will be considered below.

§4 Simulating the Population The first step is to simulate a population from which the sample will be drawn. This will consist of specifying the number of pregnant women and HIV prevalence for the n areas comprising the country. Prevalence values will be generated from a beta probability density distribution,

$$\beta_{\mu,\nu}(x) = \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)}(1-x)^{\mu-1}x^{\nu-1},\ 0<x<1, \tag{1}$$

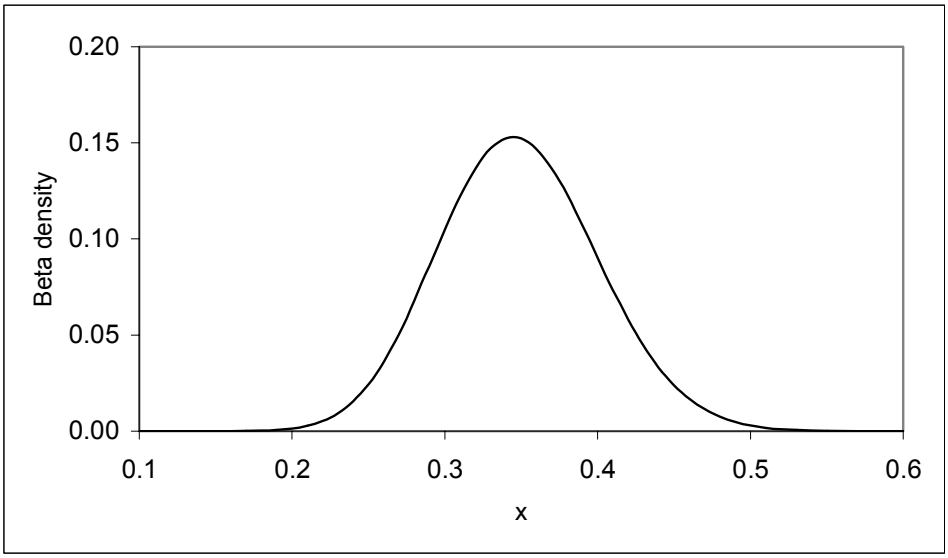
which has mean $\nu/(\nu+\mu)$ and variance $\mu\nu/[(\mu+\nu)^2(\mu+\nu+1)]$ [2]. Moment estimates are given by $\mu = \frac{m^2(1-m)}{v} - m$ and $\nu = \frac{m(1-m)^2}{v} - (1-m)$, m denoting an observed mean and v and observed variance.

1997 data for 9 sites in rural Botswana (4 values interpolated from surrounding years), for example, give the following prevalences.

Site	Prevalence
Maun	0.333
Chobe	0.383
Kweneng & K. East	0.405
Lobatse	0.337
Serowe / Pala	0.344
Tutume	0.337
Southern district	0.232
Kgatleng	0.305
Mahalapye	0.282

The mean and variance are 0.3287 and variance 0.0026, which give $\mu = 27.06$ and $\nu = 55.25$. The resulting distribution is shown in Figure 1 below.

Figure 1 Beta Density with $\mu = 27.06$ and $\nu = 55.25$



Prevalence figures for the *n* areas may be generated by the following line of **R** code

```
true.prevalence.all.areas <- rbeta(n,27.06,55.25)
```

where the desired number of areas is substituted for *n* [3]. This will return a vector of *n* values drawn from a beta distribution with the indicated parameters.

We assume for the present the same number of pregnant women in each area, so that "true" national prevalence is the simple average of the "true" prevalence over all areas.

§5 Simulating a Two Stage Sample The first stage will select *k* areas from the *n* areas comprising the country. This may be effected with the **R** command

```
true.prevalence.sample.areas <- sample(true.prevalence.all.areas,
k)
```

where `true.prevalence.all.areas` denotes the vector of prevalence values for all areas, `true.prevalence.sample.areas` the vector of values for sample areas, and *k* is replaced by the number of areas desired.

The second stage will select m_k women from the *k*-th sample area. For the present we simply sample from a binomial distribution with probability equal to the prevalence for the sampled area and number of trials equal to the number of women in the sample. The **R** command for effecting this is

```
sample.hiv.positive <- rbinom(1, nwomen,
true.prevalence.sample.area)
```

where `nwomen` denotes the number of women in the sample. This is repeated for each sample area. Sampling without replacement could be implemented with slightly more effort.

§6 Program Code The following **R** program

- takes as input (i) the vector giving true prevalence for all areas, (ii) the number of areas to be sampled, and (iii) the number of women to be sampled in each area;
- draws a sample and computes national prevalence by either the median or the binomial estimator; and
- produces as output the national prevalence estimated either by the median or the binomial estimator.

In the following `tpaa` denotes `total.prevalence.all.areas`. Lines beginning with `#` are comments. One of the estimator lines should be "commented out".

```
simulate.sample.estimate <- function(tpaa, nsites=10, nwomen=300)
{
    #stage 1: select sample areas
    tpsa <- sample(tpaa, nsites) # true prevalence for
[sample|all] areas
    #stage 2: select pregnant women from sample areas
    scsa <- rep(0,nsites) # sample counts hiv+ for sample areas
    for (i in 1:nsites) {
      scsa[i] <- rbinom(1,nwomen,tpsa[i])
     }
    #calculate estimates of national prevalence (comment out one)
    estimates <- median(scsa/nwomen)
    #estimates <- sum(scsa)/(nsites*nwomen)
    estimates
}
```

Using this program to draw samples repeatedly we generate a sampling distribution for estimated national prevalence. The **R** program code for this follows.

```
distribution.of.estimates                                          <-
function(tpaa,nsites=10,nwomen=300,nsamples=20){
    estimates <- rep(0,nsamples)
    for (i in 1:nsamples) {
        estimates[i] <- simulate.sample.estimate(tpaa, nsites, nwomen)
    }
estimates
}
```

Having defined a `true.prevalence.all.areas` vector providing the population data from which the samples are drawn (**§4**) the function is executed by

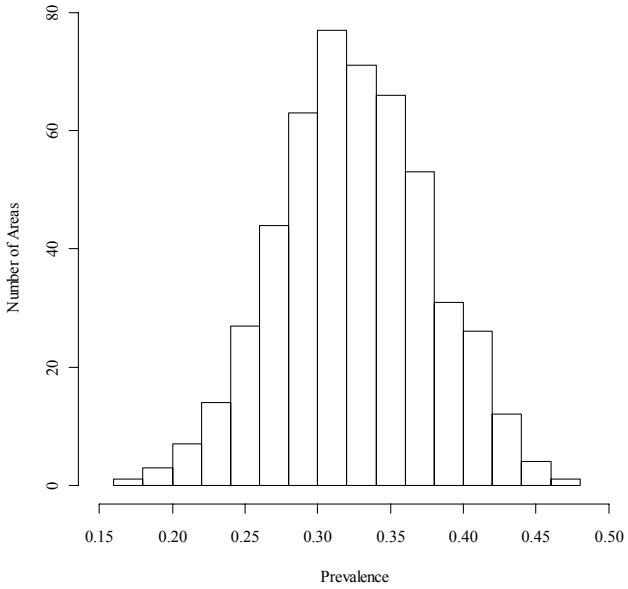
```
estimates<-
distribution.of.estimates(tpaa,nsites=10,nwomen=300,nsamples=1000)
```

The resulting vector `estimates` provides the distribution of estimates from the samples drawn.

§7 First Results "True" prevalence values for 500 areas are drawn from a beta distribution with the parameters indicated in **§3** above. True national prevalence for this population, the average of the prevalence over all 500 areas, is 0.325. The standard

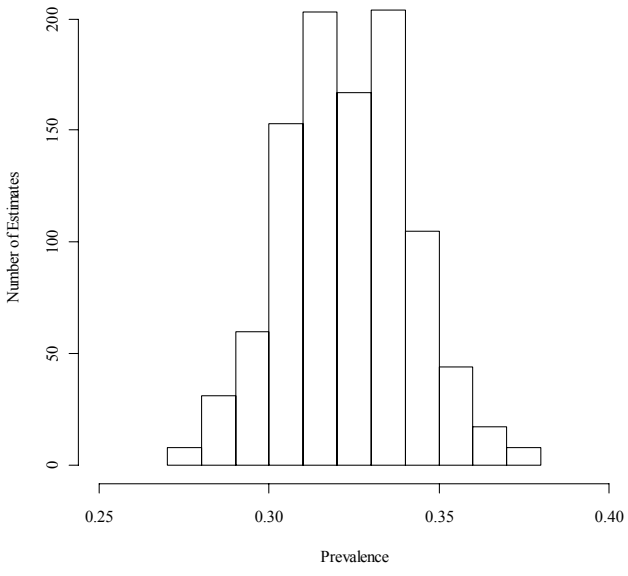
deviation over all 500 areas is 0.052. The interquartile range is 0.072. Figure 2 below shows the histogram of the distribution.

Figure 2
Distribution of True Prevalence in Population



Samples from this population will be drawn by taking a simple random sample of 10 areas and then a simple random sample of 300 women from the ANC site in each area. Drawing 1,000 such samples and estimating national prevalence as the median of the sample prevalence for the 10 sampled sites for each sample gives the sampling distribution shown in Figure 3 below.

Figure 3
Distribution of Estimates of National Prevalence
10 Areas, 300 Women, Median Estimator



The mean and median are both 0.323. The standard deviation and interquartile range are 0.019 and 0.026. The interquartile interval is thus approximately 0.323±0.026/2 = (0.310, 0.336). We expect estimates to be outside this interval about half the time.

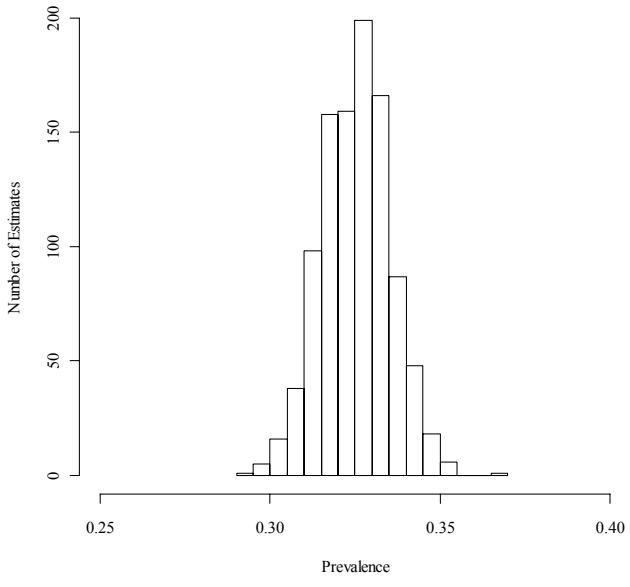
The binomial standard deviation for the same number of women (3,000) with the same probability of being HIV positive (0.323) is (root $p(1-p)/n$) 0.009. The design factor of the sample is thus approximately 0.019/0.009 = 2.1.

Recall now that these sampling errors assume perfect execution of a strictly representative two stage sample, which in turn requires the assumption that the entire country is divided into ANC catchment sized areas all served by an ANC, and finally that it is possible to take a simple random sample of women in each area sampled. Most HSS data, regarded as a sample for estimating national prevalence, fall very far short of this. Must we not conclude that actual sampling errors for national HIV prevalence estimates are *greater* than those indicated here?

§8 Sample Design Three hundred women is a reasonable sample size to obtain an estimate of prevalence for the population of women who utilize a particular ANC. From the point of view of designing a sample to provide national prevalence estimates it may be far too large. It is intuitively plausible that if women within areas are homogeneous but areas have different prevalence, better results will be produced by sampling fewer women from more areas to obtain the same total number of women. The median estimator performs very poorly in this context with small samples of women, however, so we change to the binomial estimator **(§3)**.

Sampling 30 women (rather than 300) from 100 areas (rather than 10) gives the same number of women (3,000), but a much tighter the sampling distribution, as shown in Figure 4 below.

Figure 4
Distribution of Estimates of National Prevalence
100 Areas, 30 Women, Binomial Estimator



The improvement over Figure 3—note that the horizontal scales are comparable—is dramatic. The mean here is 0.325, the standard deviation 0.010, the interquartile range 0.013.

Sample design involves cost as well as sampling error. Is it more expensive to sample 30 women from 100 sites than to sample 300 women from 10 sites? The labour of coordination and of compiling the data is greater, but the number of tests is the same and they are administered by more people (an advantage or disadvantage, depending on how much training is required). New testing methods, such as dry blood spot testing, may make samples from larger numbers of sites more feasible in the future than it is now.

§9 Discussion Based on the foregoing, a rough indication of the error in a national HIV prevalence estimate based on HSS data may be obtained in the following way, assuming that national prevalence is calculated as the median prevalence over *n* sites with 300 women tested at each site.

(1) Compute the standard deviation of the estimate on the assumption that it is based on a simple random sample, estimated by the binomial standard deviation $\sqrt{p(1-p)/n}$ where *p* is estimated prevalence and *n* is the number of women tested. (2) Multiply this value by 2 to obtain an estimate of the standard deviation of the distribution of the sampling distribution of the median estimator. (3) Multiply this value by 1.3 to estimate the interquartile range of the sampling distribution. (4) Divide the interquartile range by 2 and take this value to be the typical error, in either direction, of the estimate.

Consider for example estimated prevalence for rural Uganda, 8.4 percent, based on 24 sites [4]. Sample sizes are not available for these sites, and though they might be found in [5], for the purpose of this illustration we assume that 300 women were tested at each site, for a total sample size of 7,200. The corresponding binomial standard deviation is 0.0033. Multiplying by 2 for the sample design effect and by 1.3 to convert to interquartile range gives 0.009, half of which is 0.0045. Thus we expect the estimate of 8.4 percent may err by about 0.5 percent (absolute) in either direction.

The sample design factor of 2 applied here is based on a distribution of "true" prevalence derived from data for rural Botswana, which has a much higher prevalence and therefore a different distribution of prevalence among areas/sites. How robust is this factor against this difference? How much different would the result be if the distribution of true prevalence were re-estimated for rural Uganda?

Fitting a beta distribution to the rural Uganda data gives $\mu = 2.04$ and $\nu = 17.55$. Repeating the experiment described in **§7** with 500 areas, 24 sampled areas/sites and 300 women tested per site and drawing 1000 samples gives a standard deviation of 0.014. The corresponding binomial standard deviation (p = 0.110, n = 7,200) gives 0.004, for an estimated sample design effect of 3.5. This is considerably greater than the 2.1 initially derived (does this reflect the lower prevalence?), indicating that more work must be done to develop sample design factors for different parameter values.

§10 Conclusion The results reported here suggest that the sampling error of national HIV prevalence estimates derived from HIV Sentinel Surveillance data is *at least* 2-4 times greater than that of a simple random sample of the same number of women. Further examination of available data for the adequacy of beta distribution fits and simulation calculations to refine the results would seem to be worth pursuing.

This note has not addressed the selection problems of the HSS data, which may be more severe than sampling errors. Selection errors cannot be effectively addressed without collecting data for previously unsampled segments of the various populations. Without such data, modeling and simulation studies can do little more than reiterate their inputs. The sampling design considerations of **§8** are pertinent, however, to the design of data collection efforts.

As noted in **§1**, estimates of sampling error are pertinent to the fitting of models to observed prevalence because of the number of HSS sites available in any particular country often varies greatly from year to year. This suggests that when fitting by, e.g., least squares, *weighted* least squares fits with weights varying inversely with estimated sampling error should be used.

Acknowledgements I am grateful to Basia Zaba for numerous comments on a previous draft.

References

[1] Deming, William Edwards. 1966. *Some Theory of Sampling*. New York: Dover Publications, Inc. Reprint of 1950 edition published by John Wiley & Sons. See Chapter 5, "Multistage Sampling."

[2] Feller, William. 1966. *An Introduction to Probability Theory and Its Applications,* Volume II. New York: John Wiley & Sons, Inc.

[3] The **R** Core Team. 2000. *The R Reference Index*. Version 1.1.1 (August 15, 2000). Available online at http://www.r-project.org/ . **R** is a language and environment for statistical computing and graphics freely available at the same World Wide Web site.

[4] Reformatted data contained in **sentinelsites.zip** file provided by Basia Zaba (based on **TestData.xls** file provided by Neff Walker), **Uganda.xls** file, **input data** sheet.

[5] HIV/AIDS Surveillance Data Base. 2000. US Bureau of the Census. Available online at http://www.census.gov/ipc/www/hivaidsd.html .