Methods for modeling the HIV/AIDS epidemic in sub-Saharan Africa

Joshua A. Salomon, Emmanuela E. Gakidou, Christopher J.L. Murray

Introduction

Since the first cases of AIDS were identified in the United States nearly two decades ago, HIV/AIDS has emerged as one of the leading challenges for global public health. Particularly in sub-Saharan Africa, where the overwhelming majority of HIV and AIDS cases appear, the epidemic continues to take a massive human toll.

An understanding of the magnitude and trajectory of the HIV/AIDS epidemic, as well as the uncertainty around these parameters, is critically important both for planning and evaluating control strategies and for preparing for vaccine efficacy trials. Particularly as efforts mount to make new technologies more widely available in the developing world, tradeoffs among different potential interventions and other critical policy decisions must be based on the best possible information on the current levels and trends in the epidemic. Unfortunately, population-based epidemiological data are extremely limited in sub-Saharan Africa. Incidence data in representative study samples are rare due to the difficulty of direct measurement of population incidence and the high costs and long follow-up periods required for cohort studies. AIDS notification data represent only a fraction of new cases of AIDS and are subject to the problems of reporting delays. Information on HIV/AIDS-attributable mortality is also essential to assessments of the impact of the epidemic, but vital registration systems have extremely limited coverage in most of sub-Saharan Africa; other population-based information on mortality, while increasingly available for children through the Demographic and Health Surveys, for example, are relatively uncommon for adults.

The most widely available epidemiological data on HIV/AIDS in Africa are seroprevalence data. Population-based survey data on seroprevalence would be most desirable but only a handful of community-based surveys have been undertaken thus far [1-4]. Sentinel surveillance systems, on the other hand, have emerged in countries throughout the region, and provide information on the prevalence of infection among particular population sub-groups, including risk groups such as commercial sex workers and injecting drug users, but also including pregnant women attending antenatal clinics (ANC). In generalized epidemics as in many countries of sub-Saharan Africa, the ANC data are regarded as the closest approximation to prevalence levels in the adult population, but the exact relation between prevalence among ANC attendees and general population prevalence remains uncertain. Seroprevalence information from a range of sites globally has been compiled by the United States Bureau of the Census since 1987 [5] and made available by the Joint United Nations Programme on HIV/AIDS (UNAIDS) and the World Health Organization (WHO) in the form of Epidemiological Fact Sheets for each country.

Given the need to develop a better understanding of the levels and trends in the HIV epidemic and the limited information on which to base these estimates, the use of modeling approaches can make a valuable contribution. The goal of any modeling exercise will be to extract as much information as possible from available data in order to

provide an accurate representation of both the knowledge and uncertainty about the epidemic. Starting from seroprevalence data, a number of efforts have been made to reconstruct the past incidence of HIV infection in sub-Saharan Africa using models [6-10]

The objective of this paper is to highlight areas in which further methodological developments may be fruitful given currently available data sources. We will describe preliminary efforts at extending previous methods and suggest some avenues for further work. The primary goals of these ongoing efforts are twofold: (1) to improve the methodological basis for modeling the HIV/AIDS epidemics in sub-Saharan Africa; and (2) to produce estimates that include a meaningful characterization of at least some of the sources of uncertainty around these epidemics. The focus of this paper is on modeling HIV/AIDS epidemics in adult populations in sub-Saharan Africa. The epidemiology and impact of HIV/AIDS in children is obviously of critical importance as well, but remains outside the scope of this paper. In the following sections, we describe briefly some of the previous modeling work, introduce our preliminary work in extending these methods, present results from this exercise using the example of Zimbabwe, and discuss directions for further methodological development and refinement.

Background

A range of different types of models have been developed and applied to the estimation of HIV/AIDS epidemics in a variety of settings. Overviews of the different categories of models have been presented previously [11]. These models range in complexity from simple extrapolations of past curves [12] to complex transmission models [13-16].

One of the major traditions in modeling HIV/AIDS epidemics has been the use of backcalculation, or back-projection techniques, which produce statistical solutions to convolution equations relating the number of AIDS diagnoses over time to past trends in HIV infection and the incubation period distribution. These models were introduced more than a decade ago [17,18] and have since been applied in a host of different settings [19-26]. These models have been used almost exclusively in developed countries where AIDS notifications, while imperfect, are considerably more complete than in most developing countries. The literature on backcalculation methods includes a variety of efforts to account for different sources of uncertainty, such as the length of reporting delays and the incubation distribution [27], as well as the effects of treatment on the trajectory of the epidemic and other issues [22,28].

For modeling HIV epidemics in developing countries, the traditional backcalculation framework cannot be used in most cases, due to the paucity of reliable information on the incidence of AIDS. A modified framework was therefore developed by WHO in order to reconstruct HIV incidence curves and develop short-term projections based on the prevalence of HIV infection rather than AIDS notifications. The model developed by WHO was formalized in a software program called Epimodel [8]. Epimodel uses an input estimate of point prevalence in a reference year, combined with assumptions about HIV/AIDS progression rates and the start year of the epidemic, to reconstruct incidence curves from the beginning of the epidemic. Because there could be infinitely many different incidence curves consistent with a particular start year and point prevalence

estimate, Epimodel imposes further structure on the problem by assuming that the HIV infection rate follows a parametric curve over time based on the gamma distribution, and allowing the analyst to specify both the shape of the curve and the position on the curve in the anchor year. Thus, Epimodel may be considered a deterministic variant of the original backcalculation models.

Epimodel has been used as the basis for a series of estimates produced by the former WHO Global Programme on AIDS, and subsequently through the collaborative efforts of WHO and UNAIDS. Several sets of regional estimates have been developed since 1989 based on the estimated number of HIV-infected individuals in each region [6-9]. Prevalence of infection was estimated by country for the year 1994 [29] and then revised for 1997 to produce the first country-level epidemic estimates using Epimodel [10].

For the most recent round of WHO/UNAIDS estimates for countries in sub-Saharan Africa [10], prevalence data from antenatal clinics have been used as the starting point. The national estimates from 1994, along with more recent surveillance data from 1995 through 1997, were reviewed by expert panels in order to produce national prevalence estimates for 1997 in each country. These estimates were used as the point prevalence inputs to Epimodel, and epidemic curves were selected in each country to be consistent with the 1994 national prevalence estimates.

Given the accumulation of sentinel surveillance data over the last ten years, it is worth revisiting some of the strong restrictions in Epimodel that were necessitated by the dearth of data inputs at the time of its development. In particular, it is useful to reconsider how the statistical tools developed in the original backcalculation work might be reintroduced into models for developing countries. This is especially important given the need to characterize the uncertainty around the HIV epidemics in Africa. The preliminary work described in this paper, therefore, represents a hybrid of statistical techniques from the backcalculation work with some of the innovations developed for Epimodel. For this preliminary exercise, we have preserved a number of basic assumptions used in Epimodel but relaxed the deterministic structure imposed on the curve-fitting procedure. Using a maximum likelihood approach, we are able to use all of the available data to inform the curve-fitting exercise, as well as to represent some of the uncertainty in the estimates. Because time series surveillance data are incomplete in almost every sentinel site, we have also considered methods for augmenting the available data by imputing missing data points based on observed values and other potential covariates. As this work proceeds, further efforts are needed in critically examining various fundamental assumptions, some of which are discussed below.

Methods

The basic model used here incorporates the same change introduced in Epimodel to the original backcalculation framework, namely a shift from AIDS incidence to HIV prevalence as the starting point for the model. The model specifies that HIV prevalence at time t is a function of the number of new infections that have arisen prior to time t times the probability that individuals infected at past times have survived through time t:

$$P(t) = \int_0^t I(s)\,[1\text{-}F(t\text{-}s)]\,ds \qquad (1)$$

where $P(t)$ is prevalence at calendar time t, $I(s)$ is incidence at calendar time s, and $F(\tau)$ is the probability of dying within τ years of infection.

Beginning with this general framework, a number of different methodological issues may be examined. In this paper, we discuss one set of models used for the incidence curve $I(s)$ and incubation distribution $F(\tau)$ and a method for deriving a maximum likelihood estimate and likelihood bounds on the epidemic parameters; possible alternatives are also discussed briefly.

A discrete analog to equation (1) has been specified, and parametric forms for $I(\cdot)$ and $F(\cdot)$ have been selected to be consistent with the assumptions used in Epimodel. Specifically, a gamma distribution has been used for incidence and a Weibull distribution for the cumulative progression rate. The gamma distribution used here is described by two parameters, α and β, and has the following form:

$$I(t) = \frac{\beta^{-\alpha+1} t^{\alpha-1} \mathrm{e}^{\frac{-t}{\beta}}}{5\Gamma(\alpha)} \qquad (2)$$

Figure 1 presents 3 examples of different incidence curves that are possible using this parameterization. Some of the advantages and disadvantages of the gamma model for incidence are discussed below.

The Weibull distribution is described by two parameters, κ and ψ, and has the following form:

$$F(\tau) = 1\text{-}\exp(-\kappa\tau^{\psi}) \qquad (3)$$

The Weibull distribution has been used frequently to describe the AIDS incubation distribution [17,18,23], although the strengths and weaknesses of the model, as well as alternative assumptions, have been examined elsewhere [24].

Based on this formulation of the model, the analytical objective is to find the set of model parameters that are most likely to have produced the observed data on prevalence. Each set of parameter values generates a unique set of incidence, prevalence and mortality curves. A first round of analysis has been undertaken with the assumption that the ANC data are representative of national prevalence levels in the adult population, but the validity of this assumption is considered further below.

It would be possible to specify a likelihood function that depends only on the observed data points, and then maximize this function to identify the parameter values with the highest relative likelihood given these observations. There are several reasons, however, why it may be useful or even necessary first to address the problem of missing data values. First, in the simplest models each observation on prevalence in a particular site and particular year is considered to be independent of every other site-year observation; more sophisticated models that capture the relationships between observations within the same site or heterogeneities across the sites would require complete time series for some range of years in every site. Second, given the large number of missing values relative to observed values in many countries in sub-Saharan Africa, it may simply be impossible to compute maximum likelihood parameter estimates based only on the limited number of observations. In these cases, observed values in other sites or in other years for the same sites, as well as other covariates and information gleaned from other countries, could be helpful in predicting a distribution of likely values for the unobserved data points.

Third, if the probability that a particular value is missing is correlated with the level of the missing value, this would produce a bias that might be mitigated by filling in the gaps in the data set. Stated in other words, if the observations that are available in a particular year tend to come from sites with systematically higher or lower than average prevalence rates, then the conclusions drawn from the incomplete data set will be biased upwards or downwards accordingly.

Thus, the first step in the analysis was to apply statistical techniques for missing data imputation to the data set on prevalence in each ANC site in sub-Saharan Africa for the years 1990 to 1997. Years prior to 1990 had an insufficient number of observed values (around 50 or fewer observations per year across all 500+ sites in sub-Saharan Africa) to allow imputation of the missing values. The imputation procedure was undertaken on all ANC sites from all countries in the region simultaneously in order to borrow statistical strength from the full data set, but the subsequent backcalculation exercise was undertaken by country, and separately for the urban and non-urban sites within a country, in order to allow for the diverse epidemiological patterns that may have arisen.

A multiple imputation approach has been applied, which allows subsequent analyses to include the level of uncertainty around each imputed value, as described below. This approach has been applied frequently to survey data [30], but has also recently been used with epidemiological data on the HIV incubation distribution [31]. The statistical model used for the multiple imputation procedure was a joint multivariate normal distribution. Prevalence in each year from 1990 to 1997 was included as a separate variable in the model, along with the additional variables described below. Because the data to be imputed was prevalence data, which is bounded by 0 and 1 and distributed asymmetrically, a logit transformation was applied to the observed prevalence data prior to imputation, and the inverse transformation applied subsequently to the imputed data for use in the backcalculation model.

One of the benefits of missing data imputation procedures is that they allow potentially valuable information from variables not included in the analytical model to help predict values for the unobserved data points. In this analysis, we included a variable distinguishing between urban and non-urban sites and a variable that divides sub-Saharan Africa into 8 sub-regional country groups, to capture the expected correlations among the trajectories of HIV epidemics in neighboring countries. Additional variables, including GDP per capita, literacy rates, and primary school education rates, were also included in the imputation model. Using *Amelia*, a statistical software program designed specifically for missing data imputation [32,33], distributions for each missing value were estimated using a modified EM search algorithm, and 30 complete data sets were generated by sampling from these distributions. Across the 30 data sets, each observed value was constant, while each missing value was filled in with a random draw from the posterior distribution of that missing value. The backcalculation analysis was then performed separately on each completed data set for a country and the results combined across the 30 analyses as described below.

Although the estimation model we have specified, in its most general form, allows for maximum likelihood estimation of all four parameters in the model – two parameters of the gamma model for incidence and two parameters for the Weibull model of progression – the poor quality of the data in many instances demands a more modest estimation task.

In the work described here, we have therefore fixed the parameters on the Weibull distribution at $\kappa = 0.021$ and $\psi = 1.6$. These parameters produce a distribution with the same shape as that specified in the baseline Epimodel assumptions and a median time from HIV infection to death of 9 years. Uncertainty around the incubation distribution remains an important issue to be addressed in future work.

For each of the 30 data sets, maximum likelihood techniques were used to estimate the most likely parameter values for α and β given the observed and imputed data points on all of the sites in a country. Separate parameters were estimated for the urban and non-urban sites. Multiple simulation was then used in order to combine the results from the 30 separate analyses and to obtain bounds on the epidemic curves. For each of the 30 MLE results, 100 different sets of urban and non-urban α and β values were drawn from bivariate normal distributions with means and covariances determined by the maximum likelihood estimates. Thus, across the 30 analyses, a total of 3000 different sets of parameter values were sampled. Each set of parameter values was then used to propagate incidence curves for the urban and non-urban populations. The urban and non-urban incidence curves were combined in each of the 3000 results by weighting the two estimates in each year by the relative proportions of the national population living in urban and non-urban areas. Each of the resulting national incidence curves was then used to generate the corresponding prevalence and mortality curves based on the assumed progression rate distribution. Finally, maximum likelihood estimates and bounds on the epidemic curves were obtained by identifying the incidence, prevalence and mortality values in each year at the median, 2.5^{th} percentile and 97.5^{th} percentile of the range of 3000 sets of results.

Preliminary results using the example of Zimbabwe

Table 1 provides a summary of the data from antenatal clinics in Zimbabwe. Of the 47 sites providing at least one year of information, none has a complete series even over a subset of years, such as 1990 to 1995. Inspection of the data suggests a strong possibility that there may be potential bias introduced in analyses that do not address the problem of missing data values. For example, of the 3 sites with observations in 1996, a simple inspection of previous prevalence estimates in these sites suggests that they may deviate regularly from the prevailing levels of prevalence in the full set of sentinel sites. Two of the sites from 1996 have levels in other years that range from 1.5 to 2.5 standard deviations above the mean levels in those years, while the third site may be more representative, with levels in other years ranging from 0.2 standard deviations below to 0.7 standard deviations above the expectation in those years. Although this simple comparison is far from conclusive, it does highlight an important concern that the subset of observations in certain years may be a biased sample of the full data set if all values could be observed. It will be important to run more rigorous diagnostics on the potential biases that arise with and without imputation of missing data points. For the analysis in this paper we have used the missing data imputation algorithm described above, but comparisons to analyses on the incomplete data set will be useful in ascertaining the effects of missing data on the results of the backcalculation exercise.

Maximum likelihood parameter estimates from the 30 separate analyses have been used to generate epidemic curves and bounds on these curves as described above. Figure 2 shows the maximum likelihood estimates and likelihood bounds of the incidence, prevalence and mortality rates over time from the backcalculation models. The results indicate a dramatic rise in HIV prevalence and AIDS mortality in Zimbabwe over the last decade. It is worth noting that the likelihood bounds on these graphs reflect some but not all components of uncertainty around these estimates. Specifically, they include the uncertainty around the imputed values for missing data points given the observed values and imputation model, and the statistical uncertainty around the maximum likelihood estimates given the specified model. Notable sources of uncertainty that are omitted from the model include uncertainty around the incubation distribution, model uncertainty about the particular functional form of the incidence curve, and, as discussed below, qualitative uncertainty about the generalizability of the sentinel surveillance data.

It is important to emphasize that these results were obtained assuming that the ANC data represent prevalence levels in the entire adult population. One study from two sites in Zimbabwe has found that differences in age structure and religious affiliations of attendees to ANC sites in the study areas may lead to important selection biases in the prevalence data from these sites [34]. The results presented here must therefore be interpreted with caution. Below we discuss this question more generally within the broader regional context and consider a handful of other studies that have examined the representativeness of the ANC sentinel sites.

It is also important to seek other independent data sources that can be used to validate results based on models. One possibility is to compare the model-based mortality estimates to demographic information on AIDS-related mortality in a country, although the challenge of finding adequate population-based mortality data in sub-Saharan Africa remains daunting. Preliminary work using a series of mortality studies from Zimbabwe to validate the model results presented here suggests that these models may overestimate the level of the epidemic in Zimbabwe; considerable work remains in assessing the effects of potential biases in these mortality data on this conclusion.

Discussion

The example presented above should be viewed as a preliminary illustration of ongoing efforts to improve methods for modeling the HIV/AIDS epidemic in sub-Saharan Africa. We must note that Zimbabwe is a country in this region with comparatively rich data sources that allow the exploration of a range of different methodological refinements. The methods described here, and some of the anticipated avenues for further work, may not be applied with equal success to all countries in the region. For some countries with very little available data, such as Nigeria, efforts to apply these techniques suggest that further attention to the minimum data requirements for the method, and consideration about how to incorporate other prior information into the analyses, will be important.

We hope that this paper will stimulate critical discussions on new methods for developing epidemiological estimates given available data sources, and also help to identify priority areas for research that will enhance our understanding of the epidemic. Towards these ends, collaborative efforts are currently underway between WHO, UNAIDS, and leading

experts on HIV/AIDS, aimed at improving the empirical and methodological foundations for estimating the trends and impact of the HIV/AIDS epidemic.

Already, there are a number of important methodological and empirical questions suggested by the work described here, and we introduce only a few in the following discussion. Perhaps the most important focus relates to the representativeness of available data.

There have been a handful of studies that have addressed the question of whether prevalence rates in ANC sites are representative of the population prevalence rates. This question can be considered on three levels: (1) Do prevalence rates in ANC sites represent the general population rates in these areas among women of the same age as the ANC attendees? (2) Do prevalence rates among women in these age groups represent the overall prevalence rates in the adult population in these areas? And (3) Do prevalence rates in the sentinel areas represent national prevalence rates? The answers to these questions, of course, are likely to vary widely across different settings, due to the stage of the epidemic along with a host of other factors. The few studies that address these questions, however, provide valuable reference points for further research.

On the first question, there is evidence that women attending antenatal clinics may have similar or lower prevalence rates than women of the same ages in the general population of these areas. A study in Zambia [1] found that age-adjusted prevalence rates in women attending ANC were 24.4 to 27.5% between 1994 and 1996, compared to 31.2% in the general population in urban areas, and 12.5% compared to 17.4% in rural areas. Studies in Mwanza region, Tanzania [4,35] found that women up to age 35 attending ANCs had lower prevalence rates than women in the population of the same ages, although rates among women older than 35 years were higher in the ANC group. Another study in Kagera region, Tanzania [3] found similar results, with an overall age-adjusted prevalence of 29.4% in the general population sample, compared to 22.4% in the ANC sample.

On the second question, one of the principal concerns is whether prevalence levels are different among men and women. Berkley et al. [36] found in three studies in Uganda that women had higher prevalence rates than men, with female to male prevalence ratios of 1.42 in semi-rural communities, 1.56 in rural Rakai district and 1.31 in a national serosurvey. Standardizing on the estimated age and sex distribution in the general population, the ratios were 1.34, 1.41 and 1.19, respectively. Likewise, the Kagera study found age-adjusted prevalence rates of 29.4% among females compared to 16.7% among males. In the Zambia study cited above, age-adjusted prevalence rates were comparable in males and females in rural areas (15.4% in males and 17.4% in females) but significantly lower among males than females in urban areas (20.9% compared to 31.2%). The same findings were reported in Mwanza [35], with lower rates for males than females in urban areas (9% vs. 15%) and similar rates in non-urban areas (3% compared to 4%).

Perhaps the most critical question is whether the sentinel areas provide an adequate representation of the range of prevalence levels at the national level. Studies in various countries have found significant differences across different types of areas, reflecting important distinctions that may be missed by a broad classification of sites as urban or

non-urban. For example, one study in Arusha region, Tanzania [2] compared prevalence rates in low and high socioeconomic status urban areas, semi- urban areas, and rural areas and found that prevalence rates among women were 13.3%, 7.4%, 3.4% and 1.1% in high SES urban, low SES urban, semi-urban and rural areas, respectively, with corresponding rates among men of 5.3%, 1.0%, 0% and 2.1%. respectively. If the selection of sentinel areas in a country does not concord with the population distribution across urban, peri-urban and rural areas, then this may be an important source of selection bias in national prevalence estimates based on sentinel surveillance systems.

The problem of extrapolating from a collection of individual sites to a national prevalence level would be greatly facilitated by additional information on the different sites. For example, the catchment areas of particular hospitals would be very useful information that would allow different sites to be weighted according to the proportion of the population covered by each site. Additionally, simple geographical analyses may provide useful information on which sites are likely to have related epidemics due to proximity. If more information on the sites could be incorporated formally into epidemic models, the validity of the results would likely improve dramatically.

The use of data imputation methods to augment the incomplete data set raises another set of important methodological questions. In particular, further efforts at validating the results of imputation, and further work on identifying variables that can improve the imputation procedure are needed. Imputation at different levels of aggregation should be considered; the possibility of imputing data at the country level or sub-regional level will depend critically on the availability of additional variables that can help distinguish between different observations at these levels in order to help predict missing values.

In terms of the backcalculation model, the various assumptions described here merit further attention. One of the most critical questions is the form of the incidence curve in the model. The gamma distribution used in Epimodel and in the work described here is likely to give a poor representation of the decline in an epidemic after the peak. Other parametric forms for the incidence curve should be considered, and results from epidemic models may be consulted for qualitative insights on plausible shapes for incidence curves. It is also worth considering whether weaker parametric models, such as the splines or step functions employed by many of the practitioners of backcalculation models, are feasible in the data-poor settings of developing countries. While these weakly parametric forms will tend to reduce the bias, particularly in the most recent years of the epidemic, that may result from more strongly parametric models, too much flexibility could lead to identification problems in settings where the number of observation-years is small. The model for the incubation period distribution is another source of uncertainty that should be addressed in future work. In this area, while statistical estimation may be informative, the results from ongoing natural history cohort studies [37-40] will be most critical.

As work on modeling the HIV/AIDS epidemic proceeds, it is crucial to undertake rigorous validation exercises on the modeling methods and results. The identification of valid data sources for demographic estimates of HIV/AIDS-attributable mortality must be a priority in this area. While the statistical methods we have described here capture some of the sources of uncertainty in estimating the magnitude and trends in the HIV epidemic, there are other important sources of uncertainty that have been omitted, including

uncertainty around the model specification, progression rates and generalizability of the data used.

In spite of the level of uncertainty that remains around the epidemic, what is clear is that the HIV/AIDS epidemic is a major public health challenge that demands an effective policy response. We hope that this paper will contribute towards continuing efforts to improve understanding of the epidemic as an important step in planning and evaluating this response.

Acknowledgments

The authors gratefully acknowledge the input and critical comments of Dr. Bernhard Schwartlander, Dr. Neff Walker, and members of the UNAIDS/WHO Working Group on Global HIV/AIDS and STD Surveillance.

Reference List

1. Fylkesnes K, Ndhlovu Z, Kasumba K, Musonda RM, Sichone M. Studying dynamics of the HIV epidemic: population-based data compared with sentinel surveillance in Zambia. AIDS 1998;12(10):1227-34.

2. Mnyika KS, Klepp K-I, Kväle G, Nilssen S, Kissila PE, Ole-King'ori N. Prevalence of HIV-1 infection in urban, semi-urban and rural areas in Arusha region, Tanzania. AIDS 1994;8(10):1477-81.

3. Kwesigabo G, Killewo JZJ, Sandström A. Sentinel surveillance and cross sectional survey on HIV infection prevalence: a comparative study. East African Medical Journal 1996;73(5):298-302.

4. Kigadye R-M, Klokke A, Nicoll A, Nyamuryekung'e KM, Borgdorff M, Barongo L, Laukamm-Josten U, Lisekie F, Grosskurth H, Kigadye F. Sentinel surveillance for HIV-1 among pregnant women in a developing country: 3 years' experience and comparison with a population serosurvey. AIDS 1993;7(6):849-55.

5. United States Bureau of the Census. Recent HIV seroprevalence levels by country: February 1999. 1999; Research Note No. 26.

6. Chin J, Lwanga S, Mann JM. The global epidemiology and projected short-term demographic impact of AIDS. Population Bulletin of the United Nations 1989;27:54-68.

7. Chin J. Global estimates of AIDS cases and HIV infections: 1990. AIDS 1990;4(Suppl 1):S277-S283

8. Chin J, Lwanga SW. Estimation and projection of adult AIDS cases: a simple epidemiological model. Bulletin of the World Health Organization 1991;69(4):399-406.

9. Mertens TE, Low-Beer D. HIV and AIDS: where is the epidemic going? Bulletin of the World Health Organization 1996;74(2):121-9.

10. Joint United Nations Program on HIV/AIDS and World Health Organization. Report on the global HIV/AIDS epidemic – June 1998. 1998; UNAIDS/98.10 - WHO/EMC/VIR/98.2 - WHO/ASD/98.2.

11. Anderson RM. Mathematical and statistical studies of the epidemiology of HIV. AIDS 1989;3(6):333-46.

12. Healy MRJ, Tillett HE. Short-term extrapolation of the AIDS epidemic. Journal of the Royal Statistical Scoiety (A) 1988;151(1):50-65.

13. Anderson RM, May RM, Boily MC, Garnett GP, Rowley JT. The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS. Nature 1991;352:581-9.

14. Bongaarts J. A model of the spread of HIV infection and the demographic impact of AIDS. Statistics in Medicine 1989;8:103-20.

15. Arca M, Perucci CA, Spadea T. The epidemic dynamics of HIV-1 in Italy: modelling the interaction between intravenous drug users and heterosexual population. Statistics in Medicine 1992;11:1657-84.

16. Kault DA. The impact of sexual mixing patterns on the spread of AIDS. Mathematical Biosciences 1995;128:211-41.

17. Brookmeyer R, Gail MH. Minimum size of the acquired immunodeficiency syndrome (AIDS) epidemic in the United States. Lancet 1986;2(1320):1322

18. Gail MH, Brookmeyer R. Methods for projecting course of acquired immunodeficiency syndrome epidemic. Journal of teh National Cancer Institute 1988;80(12):900-11.

19. Brookmeyer R, Damiano A. Statistical methods for short-term projections of AIDS incidence. Statistics in Medicine 1989;8:23-34.

20. Brookmeyer R, Liao J. Statistical modelling of the AIDS epidemic for forecasting health care needs. Biometrics 1990;46:1151-63.

21. Brookmeyer R. Reconstruction and future trends of the AIDS epidemic in the United States. Science 1991;253:37-42.

22. Rosenberg PS, Gail MH, Carroll RJ. Estimating HIV prevalence and projecting AIDS incidence in the United States: a model that accounts for therapy and changes in the surveillance definition of AIDS. Statistics in Medicine 1992;11:1633-55.

23. Rosenberg PS. Backcalculation models of age-specific HIV incidence rates. Statistics in Medicine 1994;13:1975-90.

24. Marion SA, Schecter MT. Use of backcalculation for estimation of the probability of progression from HIV infection to AIDS. Statistics in Medicine 1993;12:617-31.

25. Seydel J, Kramer A, Rosenberg PS, Wittkowski KM, Gail MH. Backcalculation of the number infected with human immunodeficiency virus in Germany. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 1994;7(1):74-8.

26. Kaplan EH, Slater PE, Soskolne V. How many HIV infections are there in Israel? Reconstructing HIV incidence from AIDS case reporting. Public Health Reviews 1995;23:215-35.

27. Rosenberg PS, Gail MH, Pee D. Mean square error of estimates of HIV prevalence and short-term AIDS projections derived by backcalculation. Statistics in Medicine 1991;10:1167-80.

28. Solomon PJ, Wilson SR. Accomodating change due to treatment in the method of back projection for estimating HIV infection incidence. Biometrics 1990;46:1165-70.

29. Burton AH, Mertens TE. Provisional country estimates of prevalent adult human immunodeficiency virus infections as of end 1994: a description of the methods. International Journal of Epidemiology 1998;27:101-7.

30. Rubin DB. Multiple imputation for survey nonresponse. New York: Wiley; 1986.

31. Taylor JMG, Muñoz A, Bass SM, Saah AJ, Chmiel JS, Kingsley LA, the Multicentre AIDS Cohort Study. Estimating the distribution of times from HIV seroconversion to AIDS using multiple imputation. Statistics in Medicine 1990;9:505-14.

32. Honaker J, Joseph A, King G et al. *Amelia*: a program for missing data (Windows version). Cambridge: Harvard University; 1999; http://GKing.Harvard.edu/.

33. King G, Honaker J, Joseph A et al. Listwise deletion is evil: what to do about missing data in political science. Cambridge: Harvard University; 1998; http://GKing.Harvard.Edu/.

34. Gregson S, Zguwau T, Anderson RM, Chimbadzwa T, Chiwandiwa SK. Age and religion selection biases in HIV-1 prevalence data from antenatal clinics in Manicaland, Zimbabwe. Central African Journal of Medicine 1995;41(11):339-46.

35. Borgdorff M, Barongo L, van Jaarsveld E, Klokke A, Senkoro K, Newell J, Nicoll A, Mosha F, Grosskurth H, Swai R, et al. Sentinel surveillance for HIV-1 infection: how representative are blood donors, outpatients with fever, anaemia, or sexually transmitted disease, and antenatal clinic attenders in Mwanza Region, Tanzania. AIDS 1993;7(4):567-72.

36. Berkley S, Naamara W, Okware S, Downing R, Konde-Lule J, Wawer M, Musagaara M, Musgrave S. AIDS and infection in Uganda — are more women infected than men? AIDS 1990;4(12):1237-42.

37. Leroy V, Msellati P, Lepage P, Batungwanayo J, Hitimana D-G, Taelman H, Bogaerts J, Boineau F, Van de Perre P, Simonon A, et al. Four years of natural history of HIV-1 infection in African women: a prospective cohort study in Kigali (Rwanda), 1988-1993. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 1995;9(4):415-21.

38. Morgan D, Malamba SS, Maude GH, Okongo MJ, Wagner H-U, Mulder DW, Whitworth JA. An HIV-1 natural history cohort and survival times in rural Uganda. AIDS 1997;11(5):633-40.

39. Nunn AJ, Mulder DW, Kamali A, Ruberantwari A, Kengeya-Kayondo J-F, Whitworth J. Mortality associated with HIV-1 infection over five years in a rural Ugandan population: cohort study. BMJ 1997;315:767-71.

40. Okongo M, Morgan D, Mayanja B, Ross A, Whitworth J. Causes of death in a rural, population-based human immunodeficiency virus type 1 (HIV-1) natural history cohort in Uganda. International Journal of Epidemiology 1998;27:698-702.

Figure 1. Three examples of gamma curves.

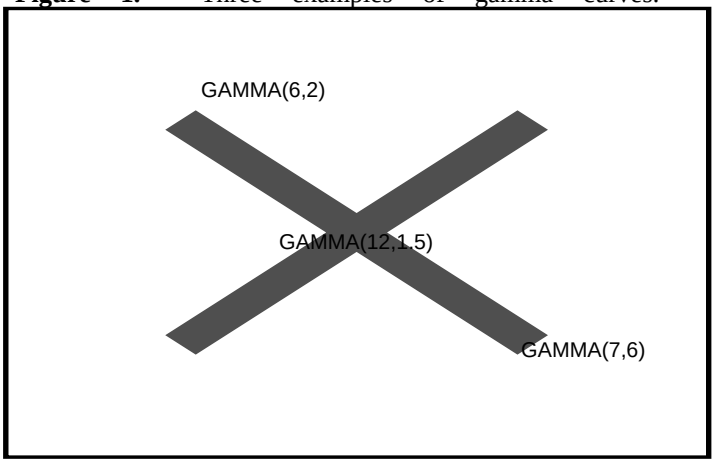

Figure 2. Incidence prevalence and mortality estimates from backcalculation models based on seroprevalence data from pregnant women in Zimbabwe.

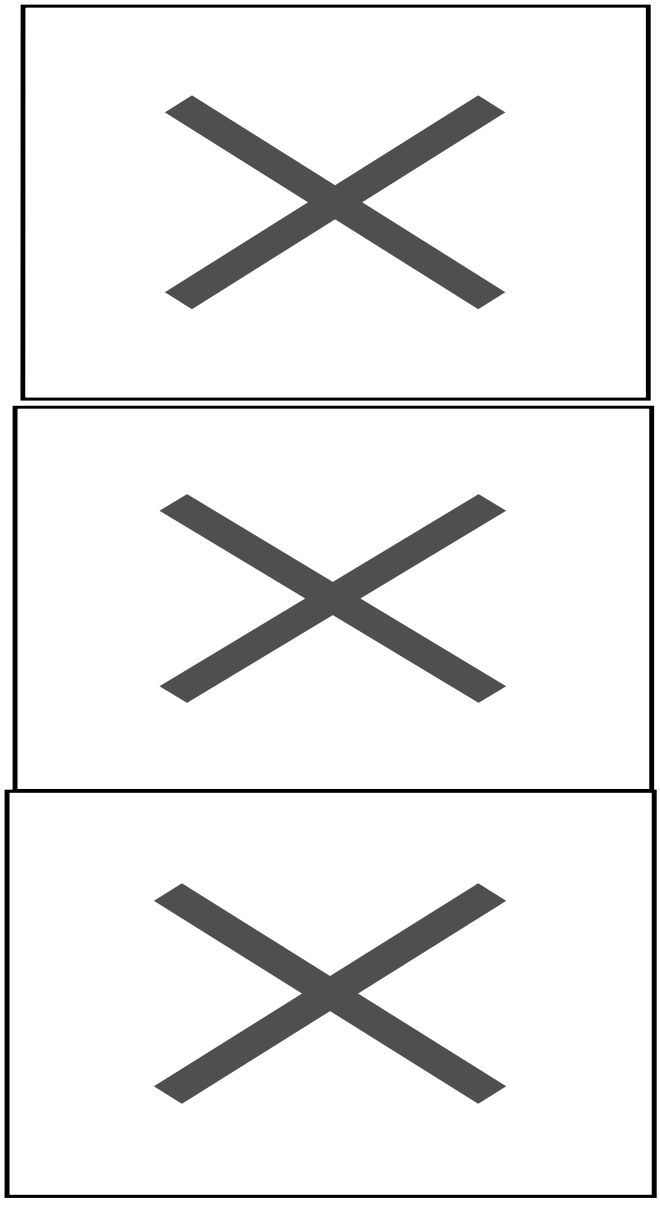

Table 1. Summary of HIV seroprevalence data from sentinel surveillance sites in antenatal clinics in Zimbabwe. Where only one site is available, the prevalence in this site is indicated as the mean, and no minimum, maximum, median or standard deviation is indicated.

	1989	1990	1991	1992	1993	1994	1995	1996
Sites in major urban areas								
number of sites	1	4	1	1	1	1	2	
minimum		16.0					30.0	
maximum		23.8					32.0	
median		18.7					31.0	
mean	10.0	19.3	17.1	29.3	25.8	30.3	31.0	
std. deviation		2.9					1.0	
Sites outside major urban areas								
number of sites		5	11	16	14	12	11	3
minimum		7.6	7.7	6.6	13.7	14.0	23.0	36.5
maximum		31.6	33.8	42.1	27.0	36.2	70.2	59.0
median		20.0	22.4	19.9	20.0	24.4	39.5	46.7
mean		21.2	20.0	20.8	20.4	24.6	40.6	47.4
std. deviation		9.2	7.5	9.0	3.8	7.1	14.8	9.2

Source: UNAIDS and WHO 1998. Epidemiological Fact Sheet on HIV/AIDS and Sexually Transmitted Diseases, Zimbabwe