MJS, Vol: 5, No. 2, July - December 2006, pp. 39-56

ISSN 0975-3311
https://doi.org/10.12725/mjs.9.5

AIDS INCUBATION PERIOD: A STATISTICAL REVIEW

Sahana Prasad, Nagaraja Rao C.** & T. Srivenkataramana****

ABSTRACT

One of the major concerns of healthcare in the world today is HIV/AIDS. The health and socioeconomic consequences of a rapid spread of AIDS are very serious. Thus we need accurate forecasts of the future course of the epidemic. The special feature of AIDS is its long incubation period, whose distribution is difficult to estimate partly due to its length and partly due to the nature of the infected cohorts being followed or identified. This article mainly discusses the features of AIDS incubation period and reviews statistical analysis of a few models developed for the estimation of the incubation period. One of the important methods of projection namely, Back Calculation method is also discussed.

Key words: *AIDS, Back Calculation, HIV, Incubation period, Transmission.*

* *Lecturer, Department of Statistics, Christ College, Bangalore-560029. Email: sahanaprasad@rediffmail.com*

** *Professor, Department of Statistics, Vijaya College, Bangalore-560004.*

*** *Professor of Statistics & Director, Bhavan- SIET Institute of Management, Bangalore-560004.*

1. Introduction

Acquired Immunodeficiency Syndrome (AIDS) is a fatal transmissible disorder of the immune system that is caused by the Human Immunodeficiency Virus (HIV). In most cases, HIV slowly attacks and destroys the immune system which is the body's defense against disease, leaving the infected individual vulnerable to malignancies and infections that eventually cause death. AIDS is the end stage of HIV infection. HIV incubation period is the time between infection and first appearance of AIDS symptoms. AIDS was recognized as a new epidemic in the year 1981 and the cause and tests to detect it were not known until 1984. It has a long incubation period. Persons who are infected by the virus may have many years of productive normal life, around 4 to 15 years, but they can infect others during this period. The prognosis for HIV infected people is bleak. At the end of the incubation period there is an increase of sickness until death occurs. The disease is found mainly in two specific age groups: children under five and adults between 15-44 years. HIV is mainly sexually transmitted, that is, it is passed on through one of the basic human activities with which people are often neither open nor comfortable. There are links between HIV and other opportunistic diseases such as TB which has further implications for public health. In general, the epidemic is still spreading and it presents a major challenge to developing countries like India,where according to UNAIDS reports, there are 5.7 million HIV positive persons.

There are four primary modes of HIV transmission: unprotected sex, infected blood transfusion, contaminated clinical needles or sharing of needles among drug abusers and from an infected mother to her new born. The risk of transmission of this virus is minimal outside the above four modes. HIV does not spread through mosquito bites, hugging, kissing, and sharing toilet facilities and same living space or by shaking hands with infected persons. The following figure (Fig.1) gives the different routes of HIV transmission among various sexual behaviour groups:

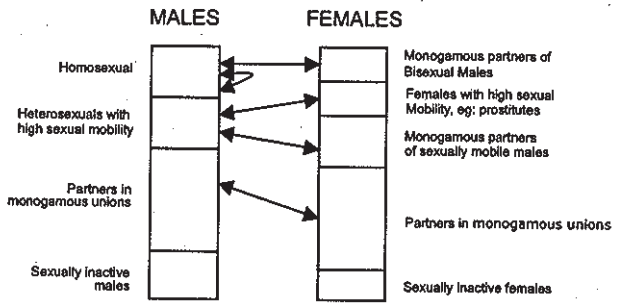


Fig.1: Principal routes for sexually transmitted HIV infection

2. Modelling of AIDS Incubation Period

Shortly after infection, HIV antibodies can be detected in the blood, whereupon an individual has *sero-converted* and is thereafter referred to as *sero-positive*. Individuals are *sero-negative* to antibodies to the AIDS virus until they are infected and become *sero-positive*. Infected individuals remain *sero-positive* i.e. prevalent with HIV infection and may eventually develop AIDS. Individuals who are *sero-positive* but free of AIDS are called *sero prevalent* and those who are initially *sero- negative* and converted to positive in a time period are *sero- incident*. *Sero-conversion* usually occurs within a few months of infection. There is more information regarding *sero-conversion* than data on infection. Therefore it is easier to estimate time from *sero- conversion* to diagnosis of AIDS or death due to AIDS.

Let X denote the chronological time of infection. For AIDS, the time between infection and sero-conversion is usually short as compared to the incubation period and accurate information on time of infection is seldom available. Thus, most statistical work assumes that all infected individuals are sero positive or that the time of infection X is the same as sero conversion. Let T denote the incubation period. The chronological time of diagnosis of AIDS is given by $Z = X + T$ which is known as the *convolution equation*. The following figure (Fig. 2) illustrates the progression of HIV to AIDS with different rates of progression.

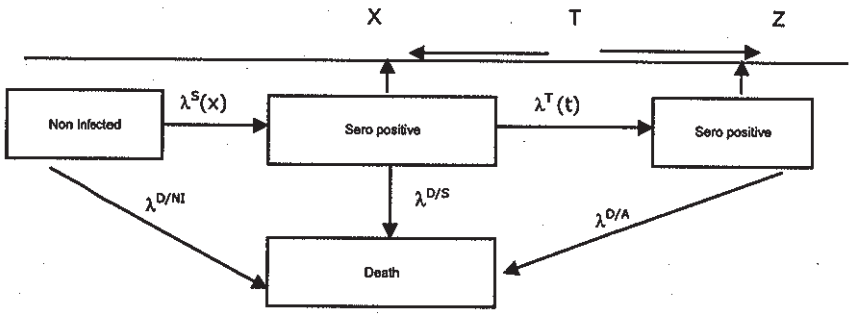


Fig. 2: Progression of HIV to AIDS

Here, $\lambda^S(x)$: Rate of sero-conversion among non infected individuals

$\lambda^T(t)$: Rate of development of AIDS among sero positive individuals

$\lambda^{D/NI}$: Rate of death among non infected persons

$\lambda^{D/S}$: Rate of death from competing causes in sero positives

$\lambda^{D/A}$: Rate of death from AIDS

A specific cohort of individuals who are longitudinally followed up is used to estimate the incubation period distribution. The cohort includes the following categories:

(i) Prevalent infected homosexual men

(ii) Individuals infected by blood transfusion

(iii) Hemophiliacs

(iv) Pediatric cases.

There is little data available on incubation period for cases associated with intravenous drug abuse and therefore estimation of incubation period for this mode is difficult. The pediatric HIV cases possess different incubation properties (Auger *et al.* 1988) than other risk groups and little work has been done so far regarding the estimation of incubation period. We now present established features of incubation period under different categories:

Prevalent cohort (unknown time of origin)

Some times, the subjects under study are already HIV infected at the time of inclusion in the study i.e. the time at first infection is unknown. This is a prevalent cohort and it consists of individuals who have a given condition such as previous infection at the time of entry and who are followed forward in time to detect the onset of disease. Individuals are at risk of infection only after onset of condition (infection) but the time of infection is an unobservable random quantity. In follow up studies of such prevalent cohorts, estimation of incubation distribution is complicated by the fact that the time of infection is usually unknown and that short incubations tend to be under sampled since individuals already diagnosed with AIDS are usually excluded in the recruitment of a prevalent cohort. Since all individuals in a prevalent cohort are censored, standard analytic tools for censored survival data do not apply. The time scales are calendar time (s), time from infection (u) and follow up time (t). Brookmeyer & Gail (1987) considered biases that arise from performing analysis on the observed follow up time scale (t) instead of the appropriate but unobservable scale of time from infection (u).

The observation period may cease prior to the onset of AIDS, so that there is a possibility of right censoring. We assume that the time of recruitment T_1 and cessation of follow up time T_2 is the same for each individual. Seropositive individuals who are already diagnosed with AIDS prior to T_1 are excluded from the cohort. The condition is therefore

$$X \leq T_1 < Z = X + T.$$

Individuals are followed until onset of AIDS or cessation of follow up, whichever is earlier. The method is illustrated with the following notations:

For the i^{th} individual, let

x_i = the calendar time of infection

$z_i = x_i + t_i$ denote the calendar time of diagnosis of AIDS

t_i = incubation time for the i^{th} individual

y_i = the calendar time of the end of follow up

δ_i = indicator which shows whether follow-up ceased due to onset of AIDS

i.e., $\delta_i = 1$ if AIDS was diagnosed at time y_i and

$= 0$ otherwise.

Since x_i is always unknown, the incubation period t_i is also unobserved even for those individuals who are followed until onset of AIDS. For an individual, prevalent at the beginning of the study T_1 and diagnosed with AIDS at time y i.e. $\delta_i = 1$, the likelihood function is given by:

$$L = \int_{T0}^{Ti} h(x)f(y-x)dx \Big/ \int_{T0}^{Ti} h(x)S(T1-x) \qquad (2.1)$$

Here, h and f are the respective densities of H and F. 'S' is the survival function associated with F. The denominator arises as prevalent infected individuals are only observed if they have not been diagnosed with AIDS by time T_1. In some cases T_1 may be sufficiently early in the epidemic and the denominator can be assumed to be one. The likelihood function is similar for an individual with a right censored observation i.e. $\delta_i = 0$ with S replacing f in the numerator.

The expression (2.1) would be simplified if all infections had occurred at a specific point in time between times T_0 and T_1. Therefore the problems encountered in prevalent cohorts arise only if the length of the interval $[T_0, T_1]$ is large which is the case for most prevalent cohorts of HIV infected homosexual men. The full likelihood function can be constructed based on products of terms as given in the above expression and the analogous terms for censored individuals. Also, some cohorts may contain seronegative individuals who seroconvert at a given time. The likelihood function can be maximized w.r.t F that is parametrically or non-parametrically described if we assume H to be known.

Non parametric estimation

Methods for estimation of incubation distribution based on prevalent cohorts assume a parametric form for both the time of infection distribution and the incubation distribution. A non parametric estimation provides a useful technique for various parametric models and to assess the goodness of fit of a parametric analysis. Since there is a lack of understanding of the disease mechanism, we cannot give a specific parametric description. A non-parametric estimation is therefore helpful to estimate the incubation period distribution. But, fully nonparametric joint estimation of the time of infection and incubation distribution in prevalent cohorts will lead to *non-identifiability* because there is no information about infection available prior to the first recruitment date. To overcome this difficulty, we can estimate a prior infection distribution from external data.

Bacchetti and Jewell (1991) proposed to use external information to estimate a sero conversion distribution. This can be used to estimate non-parametrically the distribution of incubation time from *sero-conversion* to AIDS diagnosis. This approach addresses identifiablity and makes full use of external data to provide more precise estimates of the incubation distribution. They avoided making parametric assumptions by using a discrete monthly time scale & allowing for a hazard rate for diagnosis of AIDS to be k months after sero conversion to differ for each k.

Imputation techniques

Multiple imputation is a model based technique for handling missing data problems. This technique is relevant in AIDS studies since the problem of missing data arises at different points of time in the HIV disease process. This technique can be used to estimate the distribution of times from HIV sero conversion to AIDS diagnosis. The basis of this method is to fill in the missing values to form multiple sets of complete data for further analysis. The missing values are imputed by drawing from the *predictive distribution* of the missing value given the observed data.

A cohort study of the natural history of AIDS can be thought of as consisting of two separate cohorts

(i) Sero-positive or prevalent cohort

(ii) Sero-converter or incident cohort.

As defined earlier, the sero-positive cohort consists of those subjects who already have HIV infection at enrolment and the sero-converter cohort consists of those subjects who became infected during the follow up period. One problem with the

estimation of the time to AIDS distribution from sero-positive cohorts is that infection is known to have occurred prior to a given date whereas the problems in estimating the time to AIDS distribution from sero-converter cohorts are that typically follow up times are short and the number of AIDS cases is relatively small so that accurate estimates of the distribution at long follow-up times are not attainable. Taylor *et al.* (1986) and Munoz *et al.* (1989) attempted to solve this problem by multiple imputation of the missing dates of infection and the analysis of the resulting completed data sets. Taylor *et al.* (1990) used an imputation approach in which they impute the time of AIDS diagnosis for the sero- converter cohort. They imputed events in the future (AIDS diagnoses) rather than events in the past (HIV infection). A similar approach by Moss *et al.* (1988) showed that the immunologist profile at the latest visit of participants in a cohort study was so poor that they predicted at least three quarters of the sero positive individuals in the cohort would eventually develop AIDS. The approach used by Taylor *et al.* (1990) propagated the uncertainty through the analysis in a Bayesian sense. They obtained the missing value in two stages: First a parameter value is drawn from the posterior distribution of the parameters and second, the missing value is drawn from the conditional distribution given that parameter value.

The data set used in the analysis was from a multi-centre AIDS cohort study in the U.S consisting of 4954 homo sexual or bisexual men recruited in four cities between April 1984 and March 1985. Each participant was scheduled to return at six month interval for laboratory tests, physical examination & completion of a questionnaire. They used the following seven factors: HIV antibody serology tests, ELISA and western blot test, T-helper cell percentage, platelet count, hemoglobin count and the age. The diagnosis of AIDS was not obtained from semiannual visits but rather through contact with the participant, his family, friends or physician. In defining the interval of *sero-conversion*, an HIV western blot antibody test is positive if there is any detectable antibody, however weak, and clear evidence of antibody, both ELISA and western blot, at later visits. They excluded participants with missing covariates and *sero-positive* individuals with no follow up information, as well as sero converters whose conversion interval was longer than 15 months.

Let $F(V/X, \theta)$ denote the distribution of times to AIDS measured from enrolment time for the *sero-positive* group, given covariates X and parameter q *and* $\hat{\theta}$ and cov $(\hat{\theta})$ are the MLE and their covariance obtained from the observed information matrix. A large sample is chosen so that the posterior distribution of $\hat{\theta}$ can be estimated by $N(\theta, \text{cov}(\hat{\theta}))$.

Let V denote the residual AIDS free time if $\delta = 0$ (i.e., the time from last follow up to AIDS) and $V = 0$ if $\delta = 1$ which implies $T = U + V$.

The basis of the method is to estimate the distribution of T for the sero converter group with the use of the approximately known values of U and information on the distribution of V obtained from the *sero-positive* group. The imputation technique is used to complete the data.

A value of V is drawn from the estimated F given the values of the covariates at the last visit on each *sero-converter* who had not developed AIDS. This value is added to the known value of U. Standard survival analysis techniques are used to estimate the distribution of T. The whole procedure is repeated several times and the results are combined. The parametric model used for F is an accelerated failure time model with lognormal distribution. Thus,

$$\text{Log } V_i = \theta_0 + \sum_{k=1}^{p} \theta_k X_{ik} + \sigma e_i ,$$

where V is the residual time to AIDS.

X_{ik} is the k^{th} baseline covariate measured on the i^{th} person; Θ, σ are the parameters; e has a Gaussian distribution. This model gave a better fit to the data than other models in which e was logistic gamma or Weibull. The covariates used were (*t*–helper cell percentage)^{½}, platelet count, hemoglobin count and age. They restricted attention to at most four covariates. They also found that the above combination, gave the largest maximum log likelihood. A major assumption implicit in their study was that distribution of times from infection to AIDS is constant over chronological time. Incubation period may be lengthening because of greater availability of effective treatments.

Rubin (1986) gave this standard imputation approach, which was slightly modified by Taylor *et al.* (1990).Other studies like that by Eyster *et al.*(1987), Curran *et al.* (1988) & Munoz *et al.* (1989) also gave similar results. The general result was that less than 3% of sero positives develop AIDS within 2 years of infection, about 11-32% will develop AIDS within 4 or 7 years of infection. Medley *et al* (1987), Gieseke (1988) have shown that the effect of other factors particularly age and mode of transmissions may be important in extrapolation of the incubation distribution. There is evidence of lengthening of incubation times due to greater availability of partially effective treatments.

Brookmeyer and Gail (1987) found that there are known biases in the analysis of data from *sero-positive* cohorts. With use of enrolment data as time zero when natural zero time is the unknown data of infection. They also found that while estimating the distribution of times to AIDS from a cohort study, we must consider

the statistical distribution of length biased sampling & left truncation. The key components of any model for imputation are:

i) The stochastic properties of the marker variable as disease progresses.

ii) The manner in which the hazard for AIDS onset is related to sample path of the marker.

iii) The manner in which covariates affect (i) & (ii)

Berman (1990) and Kanazawa (1991) also did some work on stochastic models for t4 helper cell counts & their use in estimating unknown times of infection but methods to investigate the relationships of covariates to the length of incubation period based on prevalent cohort data remains to be fully developed.

Interval censored data

Most statistical methods used in the analysis of survival data assume that the event that defines the start of the survival is known but the event that determines failure and hence the survival time is censored. Most of the data used in cohort studies are right censored, due to loss of follow-up, death by competing risks or study cutoff. Models for right censored data assume the origin of time scale to be known but the data of *sero-conversion* is not usually known exactly. Depending on the study design and entry criteria, a cohort study may contain prospectively and retrospectively identified *sero-converters* as well as *sero-prevalent* cases that can be defined as under:

i) *Prospectively identified sero-converters* have a last *sero-negative* and first *sero-positive* test result both obtained during follow up.

ii) *Retrospectively identified sero-converters* are *sero-positive* at entry but they have an earlier date at which they were sero negative for example, through a test result obtained from stored blood samples.

iii) *Sero-prevalent* cases are already *sero-positive* at their date of entry into the study.

In addition, there may be persons who are *sero-negative* until the end of follow up. Thus the date of *sero-conversion* can be left censored, interval censored or right censored but is hardly ever observed exactly. The time between infection & *sero-conversion* appears to be quite short (1-3 months) compared with the latency between infection & AIDS.

Doubly Censored Data

This refers to the data in which the initiating and terminating events that determine the incubation period are censored in the same individual. An example could be of hemophiliacs who became infected with HIV because the blood factor they received was contaminated with HIV. Since the blood samples from these individuals are periodically collected or stored, they can be retrospectively tested to determine a time interval during which the infection occurred. Since the incubation period between infection with HIV & development of AIDS can be very long, many of the hemophiliacs infected this way still have not developed AIDS.

Geskus (2001) considered five different methods for doubly censored data. They are:

(a) *MID- midpoint imputation* in which we impute the mid point between last negative and positive test as the date of seroconversion. The incubation period is estimated via methods for the right censored data.

(b) *EXP: Conditional mean imputation* in which the expected date of seroconversion, based on G (where G refers to the estimate of the seroconversion distribution) conditionally on date of last negative and first positive test. Incubation time is estimated via methods for right censored data.

(c) *RAN: Multiple imputation* in which the date of seroconversion is imputed based on a random draw from G and conditionally on his date of last negative and first positive test. Incubation period is estimated via methods for right censored data. This procedure is repeated several times and the average of results is considered.

(d) *LIK: Maximum likelihood* with fixed seroconversion distribution in which one maximizes the likelihood in F with G fixed.

(e) *FULL: Full likelihood* procedure in which the likelihood is maximized with respect to both F and G simultaneously.

The methods *MID, EXP, RAN* are all based on imputation of a date of seroconversion after which methods for right censored data can be used for estimation of F.

These methods were applied to data from the Amsterdam cohort studies on injection drug users and the results were compared. Survival estimates diverged to some extent. They concluded that conditional mean imputation gives almost unbiased estimates and multiple imputation yields estimates which are biased downwards. *FULL* does not yield better results than *EXP* with respect to the incubation time estimate although the estimate of seroconversion distribution was better in this approach.

3. Stage Specific Markov Model

Data available on the progression of individuals to AIDS are generally arranged in cohorts of persons who were recently infected or whose serum specifications were found to be positive (Jaffe *et al*; 1985). Kay (1986) fitted a staged stochastic model to such cohorts with use of Maximum likelihood methods & then used the fitted models to estimate the probability density function of AIDS incubation period. Persons who are infected with HIV progress through various stages of infection (Redfield *et al*. 1986, Seligman *et al*. (1987), Hethcote, (1987, 1989).Longini *et al*. (1989, 1990) gave a natural model for this progression by a sequence of five phases:

i) *Pre-antibody*: a person is infected but not antibody sero positive some persons have acute illness in this phase.

ii) *Antibody*: sero positive but asymptomatic

iii) *Symptomatic*: a person develops an abnormal hematologic indicator

iv) *Clinical AIDS*

v) *Death due to AIDS*

The progress towards AIDS coincides with a decline in the number of CD4 lymphocytes and thus CD4 cell count intervals can be used as stages and the progression through these stages can be measured. The authors have used a continuous time Markov model to model the decline of CD4 cells in HIV-infected persons. The following table presents the estimated values of transition rates (γ_i) along with the mean waiting times in each stage μ.

Estimated values of γ and mean waiting time μ in each stage of infection with no cofactors.

Stage i	CD4 cells count interval	Transition Rate γ_i in months	Mean waiting time $\hat{\mu}_i$ in years	Cumulative waiting time in years
1	> 899	0.0764	1.1	1.1
2	700-899	0.0665	1.3	2.4
3	500-699	0.0499	1.7	4.1
4	350-499	0.0429	1.9	6.0
5	200-349	0.0408	2.0	8.0
6	1-199	0.0529	1.6	9.6

The transition rates between the stages 1-6 were estimated from data on 1796 HIV – positive individuals in United States army using a continuous time Markov model. They used a persistence criteria that two consecutive measurements are needed to confirm a true reduction in CD4-cell count. This is because CD4-cell levels in individuals can vary due to measurement error & changing physiological conditions. The estimated mean time from sero conversion to when the CD4 cell count is persistently below 500 is 4.1 years, the mean time until it is below 200 is 8.0 yrs and the mean waiting time from sero conversion to AIDS diagnosis is 9.6 yrs. The data were also analyzed for three age groups ≤ 25, 26-30 and > 30. The progression rates were the same for CD4 cell counts ≥ 500 and the two older groups progressed faster when CD4 cell counts < 500. The above authors modeled the progression of an infected individual through the stages of infection and ultimately to death as a time-homogeneous Markov process in which the stages (i) – (iv) are transient and stage (v) is an absorbing state.

4. Application of Incubation Period in Back Calculation

Back Calculation (BC) method is a method in which the number of AIDS cases can be projected from those already infected with the AIDS virus. This projected number can be considered as the lower bound as this number will be expected even if there are no future infections. It presupposes the knowledge of the incubation distribution among the infected that can develop AIDS. No assumptions are required about the number of infected individuals or the probability of an infected individual eventually developing AIDS. This method does not account for further infection cases but can produce accurate short-term projection because of the long incubation period. In this procedure, we calculate 'back' from AIDS incidence data to the numbers previously infected. Brookmeyer & Gail found that the short-term projections are not very sensitive to assumed incubation distribution and the parametric model for density of infection times as long term projections. This method is illustrated briefly as follows:

The BC method depends on 3 key components:

(1) *HIV infection density*

(2) *Incubation distribution*

(3) *Observed counts of AIDS cases over time.*

The method can be described using the following notations:

T_0: Beginning time of the epidemic

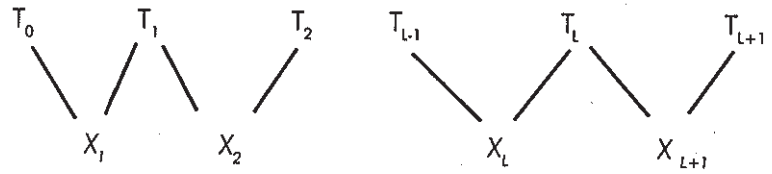
$T_0 T_1 \ldots\ldots T_L$: Calendar dates

$(T_{i-1}, T_i)\ i=1 \ldots L$ Non overlapping intervals of time

X_i: Number of AIDS cases diagnosed in the jth interval

X_{L+1}: Number infected before T_L but not yet diagnosed.

Schematic representation of available data ($X_1 \ldots X_L$) used to estimate X_{L+1} is the following:



Let N denote the total number diagnosed upto the year $L+1$. Then,

$$N=\sum_{i=1}^{L+1} X_i .$$

X_1 to X_L are known from the records whereas X_{L+1} is to be projected. The vector of counts $X=[X_1, X_2 \ldots, X_L X_{L+1}]$ has a multinomial distribution with unknown sample size N and cell probabilities i.e. $P= [P_1, P_2 \ldots, P_L, 1-P^*]$ where $P^*=\acute{O}P_i$.

If P_i is the probability that a susceptible individual infected before year T_L is diagnosed in the *jth* interval, then, $P_i = \int I(s)\ [F\ (T_i-s\ /s)-F\ (T_{i-1}-s/s)\ ds]$, where $F\ (.)$ denotes the incubation distribution i.e. $F\ (t)$ =Pr (AIDS diagnosis occurs within time t). This distribution depends on the time of infection S and $F\ (t/s) =0$ for all $t \leq 0$ and all S. $I(s)$ is the infection curve which represents the probability density function of the infection at time S for N individuals. The external information on $F\ (.)$ together with the observed $[X_1, X_2 \ldots, X_L]$ is used to estimate $I\ (.)$. The density $I(s)$ is assumed to come from a parametric family with unknown parameter θ. The multinomial likelihood is maximized to obtain joint estimates of N and θ. Dempster *et al* (1977) used EM algorithm to obtain these estimates. A short-term projections of cumulative AIDS incidence up to the future year T_{L+1} is then given by

$$\Sigma Xj + N \int I(s;q)\ [F\ (T_{L+1}-s)-F\ (T_L-s)]\ ds.$$

This is an MLE and gives lower bound for the future AIDS counts under the assumption that no new infection occurs after T_L.

The discrete version of back-projection method given by Chau *et al* is as follows:

Let $t=1,2,\ldots,t$ be the time units for the data. It is assumed that before $t=1$ the disease has not yet appeared and t is the latest time where reliable data can be obtained. In the analysis t is then taken to be 22(i.e the year 2000). Any convenient time unit, such as a year or quarter, can be adopted. In this paper, the authors used yearly numbers.

The yearly mean incidence of AIDS μ_t can be expressed in terms of the yearly mean incidence of HIV λ_1 by the convolution equation

$$\mu_t = \sum_{s=1}^{t} \lambda_s f_{t-s,s}$$

where $f_{d,s}$ is the probability function of incubation period of length d starting in the years:

Using the EM algorithm, estimes of the HIV incidence are updated by

$$\lambda_t^{[L+1]} = (\lambda_t^{L} / F_{\tau-t,t}) \; / \sum_{d=0}^{\tau-1} a_{t+d} f_{d,t} \; / \sum_{i=1}^{t+d} \lambda_t^{[L]} f_{t+d-i,i}$$

where $F_{t-i,t} = \sum_{d=0}^{\tau-t} f_{d,t}$ and a_t denotes the observed number of diagnosis AIDS in time unit t. A smoothing step is applied to give the updated estimated $\lambda_1^{[L+1]}$, $t=1,2,\ldots,t$.

A modified back-projection method using HIV diagnosis data employs the same mechanism as the original method, except that μ_t', the yearly mean number of HIV positive diagnoses are use in place of μ_t in the equation. Let D' be the incubation period diagnosis and $f'_{d,s}$ be the corresponding probability function. Therefore, the modified convolution equation becomes

$$\mu_t' = \sum_{s=1}^{t} \lambda_s f'_{t-s,s}$$

As suggested by Cui and Becker (2000), the hazard function for HIV diagnosis can be expressed as an additive model through the hazard function for AIDS, $\rho(x/u)$. It is assumed that there are two sources of hazard for the individual to have HIV

positive test, performed because of illness, whereas routine tests are tests performed otherwise, regardless of health status. The hazard function r'(x/u) of an individual infected in the year u for having a positive HIV test is

$$\rho'(x/u)=\nu+\gamma\rho(x/u) \text{ if } x+u\geq\tau_0 \text{ and } 0 \text{ otherwise,}$$

where t_0 is the time when HIV diagnostic tests became available and n denotes the constant hazard from routine tests faced by an individual. The symptom-related testing is assumed to be proportional to the AIDS diagnosis hazard through the proportional coefficient γ. Parameters ν and γ are assumed to be known from other studies. Then the EMS algorithm is applied in the same manner as the original method, with a window of width 3 for the smoothing step. The corresponding weights are given by $w_o=w_2=0.1$ and $w_1=0.8$. An adjustment is also made to allow for the zero hazard in the early years.

5. Discussion

Estimation of the incubation period of AIDS is an important field of research in the studies of HIV/AIDS, which is considered to be a global health problem. There is an urgent need for more accurate forecasts for the future course of the epidemic. The incubation period of AIDS is a unique feature of this disease, which can be modeled using various statistical techniques. There is a change in the length of incubation period in the past decade due to the effect of retroviral treatments. This paper focuses on the incubation period for prevalent cohorts, for transfusion-associated cases and for censored data. Of these, the estimation for prevalent cohort and for censored data is of unknown time origin and transfusion associated cases are retrospectively ascertained. Statistical models, either parametric or nonparametric, can estimate the incubation period distribution more efficiently. Parametric estimates have usually used a Weibull or a gamma distribution. Data from prospective study of seroconverters clearly show an increasing hazard for several years following infection. Rates of progression to AIDS are very low in the first 2 years after infection and increase thereafter. The Back Calculation method is a very important method for projecting the size of the epidemic. The median time from HIV to AIDS is about ten years in developing countries. No estimates of incubation period for Indian data are available. Estimation of incubation period for different modes of spread is also not available at the present. Further, there are some difficulties in applying the method of Back Calculation to Indian data. Incubation period and its distribution play a key role in most of the statistical studies of HIV/AIDS. Thus, their accurate estimation is very important. In cases of prevalent data where the time of infection is unknown, non-parametric and

imputation techniques are suitable. For interval censored data, various imputation techniques have been suggested. For transfusion associated cases, it has been found that the non-parametric approach has been more appropriate. A natural model for progression to AIDS can be based on t4 cell counts as these counts form an important marker for disease progression. Also, since the spread mechanism of the virus in the four modes of transmission is different, there is a possibility that the incubation period may vary from one mode of spread to another. Once the estimation of incubation period is available mode-wise, BC can be adopted for different modes separately for accurate projection of aids counts. These estimates in turn help adequate planning for providing health care to the needy.

References

1. Auger I, Thomas P, De Gruttola V, *et al. Incubation periods for pediatric AIDS patients.* Nature 1988; 336:575-577.

2. Bacchetti Peter (1990). *Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns.* Journal of the American Statistical Association, 85, 1002-1009.

3. Bacchetti Peter and Andrew Moss (1989). *Incubation period of AIDS in San Francisco,* Nature, 338,251-253.

4. Bacchetti Peter and Nicholas Jewell (1996). *Non parametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times,* Biometrics, 47, 947-960.

5. Bacchetti P, Moss AR (1989). *Incubation period of AIDS in San Francisco.* Nature; 338, 251-253.

6. Bailey. N.T.J (1975). *Mathematical theory of infectious diseases and its applications.* Hafner Press, New York.

7. Becker N.G (1998). *Method of back-projection.* Encyclopedia of Statistical Sciences, 2, 43-46.

8. Becker N.G., James J.C. Lewis, Zhengfeng Li, and Ann McDonald (2003). *Age-specific back-projection if HIV diagnosis data,* Statistics in Medicine, 22, 2177-2190.

9. Bellocco Rino and Ian .C. Marschner (1999). *Joint analysis of HIV and AIDS surveillance data in back calculation,* Statistics in Medicine, 19, 297-311.

10. Bongaarts John (1989). *A Model of the spread of HIV infection, and the demographic impact of AIDS,* Statistics in Medicine, 8, 103-120.

11. Brookmeyer R, Gail (1994). *AIDS Epidemiology: A quantitative approach.* Oxford University Press.

12. Brookmeyer R, Goedert JJ (1989). *Censoring in an epidemic with an application to hemophilia-associated AIDS.* Biometrics; 45:325-335.

13. Brookmeyer and Quinn (1995). *Estimation of current HIV incidence rates from a cross-sectional survey using early diagnostic tests*. Am J Epidemiol, 141, 166-172.

14. Brookmeyer Ron, and Mitchell .H. Gail (1988). *Method for obtaining short term projections and lower bounds on the size of the AIDS Epidemic*, Journal of the American Statistical Association, 83,301-309.

15. P.H Chau, Paul S. F. Yip and Jisheng S. Cui (2003). *Reconstructing the incidence of human immunodeficiency virus (HIV) in Hong Kong by using data from HIV posiyive tests and diagnosis of acquired immune deficiency syndrome*. Appl. Statist., 52, 237-248.

16. Dempster, A.P., Laird, N.M., and Rubin, D.B (1977). *'Maximum Likelihood from incomplete Data via the EM Algorithm*. Journal of the Royal Statistical society-B, 39, 1-22.

17. Enger C, Graham N, Peng Y, et al (1996). *Survival from early, intermediate, and late stages of HIV infection*. JAMA; 275:1329-1334.

18. Foulks Mary A. (1998). *Advances in HIV/AIDS Statistical Methodology over the past decade*, Statistics in Medicine, 17, 1-25.

19. Gail MH, Tan WY, Pee D, et al (1997). *Survival after AIDS diagnosis in a cohort of hemophilia patients*. Multicenter Hemophilia Cohort Study. J Acquir Immune Defic Syndr Hum Retrovirol; 15:363-369.

20. Geskus B. Ronald (2000). *On the inclusion of prevalent cases in HIV/AIDS natural history studies through a marker-based estimate of time since seroconversion*, Statistics in Medicine, 19,1753-1769.

21. Gigli Anna and Arduino Verdecchia (1999). *Uncertainty of AIDS incubation time and its effects on back-calculation estimates*, Statistics in Medicine, 19, 175-189.

22. Jewell P. Nicholas (1990). *Some statistical issues of the Epidemiology of AIDS*, Statistics in Medicine, 9, 1387-1416.

23. Kalbfleisch J.D and J.F.Lawless (1988). *Estimating the incubation period for AIDS patients*, Nature, 339, 504-505.

24. Lagakos S, De Gruttola V (1989). *The conditional latency distribution of AIDS for persons infected by blood transfusion*. J Acquir Immune Defic Syndr Hum Retrovirol; 2:84-87.

25. Longini M. Ira Jr, W. Scott Clark, Robert H. Byers, John W. Ward, William W. Darrow, George F. Lemp and Herbert Hethcote (1989). *Statistical Analysis of the Stages of HIV infection using a Markov model*. Statistics in Medicine, 8, 831-843.

26. Longini M. Ira Jr, Robert H. Byers, Nancy A. Hessol, and Wai .Y. Tan (1991). *Estimating the stage-specific numbers of HIV infection using a Markov Model and Back-Calculation*, Statistics in Medicine, 11, 831-843.

27. Longini Jr IM, Clark WS, Gardner LI, et al (1991). *The dynamics of CD4+ T-lymphocyte decline in HIV-infected individuals: A Markov modeling approach*. J Acquir Immune Defic Syndr Hum Retrovirol; 4:1141-1147.

28. Lui et al (1988). *A model based estimate of the mean incubation period for AIDS in homosexual men*. Science, 240, 1333-1335.

29. Mariotto B. Angela and Arduino Verdecchia (2000). *Using AIDS mortality data to reconstruct HIV/AIDS epidemics*, Statistics in Medicine, 19, 161-174.

30. May, R.M. and Anderson, R.M (1986). *'Transmission of dynamics of HIV infection'*, Nature (London), 326, 137-142.

31. Medley G, Anderson R, Cox D, et al (1987). *Incubation period of AIDS in patients infected via blood transfusion*. Nature; 328:719-721.

32. Moss AR, Bacchetti P, Osmond D, et al (1988). *Seropositivity for HIV and the development of AIDS or AIDS related condition: Three year follow up of the San Francisco General Hospital cohort*. BMJ; 296:745-750.

33. Rao Nagaraja C. and Srivenkataramana T. (.2001). *Projection of HIV infections in India- An alternative to back-calculation*. Current Science, 81, 1302-1307.

34. Rao Nagaraja C (1994). *Recent developments in randomized response techniques (M.Phil dissertation)*.

35. Taylor, J.M.G., Schartz, K., and Detels, R (1986). *The time from infection with HIV virus to the onset of AIDS*. Journal of infectious Diseases, 154, 694-697.

36. Zeger L. Scott and Lai-Chu See, Peter J. Diggle (1989). *Statistical Methods for monitoring the AIDS epidemic*, Statistics in Medicine, 8, 3-21.

37. Ziger S.L and Diggle P.J (1989). *Statistics in Medicine*, 8.