A STOCHASTIC POINT PROCESS MODEL OF THE INCUBATION PERIOD OF A HIV INFECTED INDIVIDUAL

Venkata S.S. Yadavalli
Department of Industrial and Systems Engineering, University of Pretoria, Pretoria, South Africa
E-mail: *yadavali@postino.up.ac.za*, Tel. no.: +27-826037949, Fax: +27-12-3625103
Moremi Morire OreOluwapo Labeodan
Department of Statistics, University of Pretoria, Pretoria, South Africa
E-mail: orelabeodan2001@yahoo.co.uk, Tel. No. +27-828259129
Swaminathan Udayabaskaran
Department of Computer Applications, Vel Tech Dr. RR & Dr. SR Technical University, Avadi, Chennai, India
E-mail: *s_udayabaskaran@hotmail.com*

Key words: Distribution function; HIV; incubation period; seroconversion time; stochastic model.

Summary: The inability to get people to regularly test and know their HIV status has caused the widespread unavailability of correct and comprehensive data on HIV infection especially the time at which an individual was first infected. Hence, mathematical scientists have relied extensively on inference obtained from small samples to estimate the HIV incubation and seroconversion times. We set out to obtain in this paper, (i) the distribution functions of the HIV incubation period and seroconversion time by considering the stochastic behaviours of the members of the population under discussion, and (ii) the method of estimation that gives the best parameter estimate by building on previous work of Lui et al. (1988) and Medley et al. (1988). We obtained a one-parameter family distribution for the incubation period and a two-parameter family distribution for the seroconversion time. Data on homosexual individuals were used since we built on past work of Lui

AMS: 60G55

et al. (1988). Also AIDS incidence projection was done using the back-calculation method. However, the shortfall of the back-calculation method was not addressed in this paper as this is meant for further research.

1. Introduction

Acquired Immune Deficiency Syndrome (AIDS) is a fatal but containable disease caused by the retrovirus HIV. It is found that there is a risk of contracting HIV infection from exposure to infected persons. The exposure can be through the sharing of intravenous hypodermic needles with infected persons, transfusion of HIV infected blood, mother-to-child transmission at birth, or performing a sexual act with HIV infected persons. As sex plays an important role in human life, the virus has the ability to be quickly transmitted from one infected individual to either an infected or non-infected individual by the pattern of their intimate behaviour. Since the behaviour is highly stochastic, the time for a susceptible to become an infective is unpredictable. Hence, the dynamics of the spread of HIV presents several perplexing difficulties even in the case of a specific community such as a population of transfusion related cases of AIDS (Medley et al. 1988). The foremost difficulty that baffles model builders is the incubation period of HIV. The incubation period (**IT**) of HIV in an infected individual is the period from the time of infection to the time of the first diagnosis of an opportunistic disease associated with AIDS. According to Medley et al. (1988), one of the striking features of AIDS is that the incubation period appears to be both long and highly variable. Usually, the time of infection is not known in several cases. However, the seroconversion time (**ST**) (i.e., the time at which an infected individual becomes HIV positive) may be known in many cases. The latent

period, namely, the interval between the time of infection and the time of seroconversion is small (in weeks) compared to the incubation period (in years) of HIV. Hence, the time of infection is taken to be the time of seroconversion.

Studies on the estimation of the HIV incubation period have been carried out in the past. For instance, Medley et al. (1988) in their study observed that the data on the time of infection was incomplete and estimated the mean incubation period to be 4.5 years to 15 years. Chevret et al. (1992) developed a new approach for estimating the incubation period of acquired immunodeficiency syndrome (AIDS) based on age distributions. They expressed the incubation period as the difference between the age at time of diagnosis and the age at time of contamination. By assuming independence between age at time of infection and incubation period, the age distribution of newly diagnosed AIDS cases was given as the convolution product between the distributions of the age of freshly infected patients and of the incubation times. Hence, AIDS incubation time could therefore be estimated from the age distribution of newly HIV infected subjects and newly diagnosed AIDS cases.

Lee (1999) estimated the maturity of the HIV infection and the incubation period of AIDS by using data from 363 seroprevalent (i.e. those who were AIDS free at entry) Korean AIDS patients (including 59 seroincident cases). He proposed two methods for computing the unknown times since seroconversion: (a) fitting Weibull regression with the marker of matured CD4+T cell count for seroincident cohorts, and (b), using a random effects model with CD4+T cell count as a response for repeated measures from which the times since seroconversion can inversely be extracted.

Rao and Kakehashi (2005) estimated HIV incidence density from prevalence data and also the incubation time distribution by using the

deconvolution technique and maximum likelihood method to estimate parameters. The difference was that their data was not based on homosexual men/women.

Several mathematical and statistical analyses have been proposed in the recent past to assimilate the data and provide information about the dynamics of the epidemic (Anderson and May, 1991). In the statistical analyses of the data, the gamma, Gompertz, Lognormal, Normal and Weibull distributions were used to model the distribution function $F(t)$ of the incubation period (Brookmeyer and Gail, 1994; Anbupalam et al., 2002). The advantages and disadvantages of using each of these models are outlined in Brookmeyer and Gail (1994). In particular, the Weibull model is used in situations where it is hypothesized that the hazard function $\lambda(t)$ increases indefinitely and is proportional to a power of time from infection (Brookmeyer and Gail, 1994). The hazard function quantifies how the risk of AIDS evolves with time from infection and is given by

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{1}$$

where $f(t) = F'(t)$ and $S(t) = 1 - F(t)$ are the probability density function (p.d.f.) and the survival function (s.f.) of the incubation period respectively. However, as Brookmeyer and Gail (1994) have pointed out, the hazard function $\lambda(t)$ should be consistent with epidemiological data and with theoretical considerations of the pathogenesis of HIV infection. Not much attention has been paid to the formulation of the distribution functions (hence the hazard functions) of the latent and the incubation periods by considering the stochastic behavioural aspects of the members of the population under study. Accordingly, in this paper, two stochastic models are presented namely:

(i) Model I which is devoted to the determination of the distribution function of the time to infection (i.e. the time period from the entry of a susceptible in the specified community till he/she tested HIV

positive) of a susceptible.

(ii) Model II which determines the distribution function for the incubation period (i.e., the period from the time of seroconversion till the onset of overt symptom of AIDS).

Essentially, a two-parameter family distribution function for the time to infection and a one-parameter family of distribution for the incubation period are obtained. It is observed that the distribution function of the incubation period serves as a good fit for the data provided by Lui et al. (1988). Further, the distribution function is used to project AIDS incidence by back-calculation (Brookmeyer and Gail, 1994).

The lay-out of this paper is as follows: In Section 2, a stochastic model for the determination of the p.d.f., $q(t),$ of the time interval ST between the time of entry of an individual into a population of homosexuals and the time of his/her seroconversion (becoming HIV positive) is proposed. In Section 2.1 a two-parameter family of the probability distribution function of ST is obtained. The moments of ST are obtained in Section 2.2 and the problem of estimation of the parameters of $q(t)$ is considered in Section 2.3. In Section 3, a stochastic model for the determination of the probability function, $p_n,$ of the incubation period (IT) is proposed. A one-parameter family of the probability function p_n of IT is obtained in Section 3.1 while the moments of IT are obtained in Section 3.2. The problem of estimation of the parameter of p_n is considered in Section 3.3 and illustrated by a numerical example in Section 3.3.4. The method of back-calculation is used in Section 4 to obtain AIDS projection for a set of sample data.

2. A stochastic model for the time to infection

Consider a population of homosexual individuals consisting of susceptibles and infectives. Assume that at time $t = 0,$ a new member who is tested HIV

negative enters into the population and makes sexual contacts with members of the population. Assume further that his/her contacts occur at random time points which follow a Poisson process with parameter $\lambda,\ \lambda > 0$. Let the probability that the individual who has already had n contacts up to time t when he/she tested HIV positive for the first time in the interval $(t, t+\triangle)$ be given by

$$n\mu\triangle + o(\triangle),\ \mu > 0. \tag{2}$$

Let the time to infection of the individual be represented by the random variable ST. In the next section, we obtain the p.d.f. of ST.

2.1 The probability distribution function of the time to infection

We define the p.d.f. of ST by

$$q(t) = \lim_{\Delta \to 0} \frac{\Pr\{t < ST < t+\triangle\}}{\triangle}. \tag{3}$$

Then $q(t)\triangle$ represents the probability that the individual tests HIV positive for the first time in the interval $(t, t+\triangle)$. At least one contact is needed to get infected with HIV and so the infection occurs either in the first contact or in a subsequent contact. Hence, we obtain

$$\begin{aligned} q(t) &= e^{-\lambda t}\lambda \copyright e^{-(\lambda+\mu)t}\mu + e^{-\lambda t}\lambda \copyright \\ &\quad \times \sum_{n=2}^{\infty} e^{-(\lambda+\mu)t}\lambda \copyright \ldots \copyright e^{-(\lambda+(n-1))}\lambda \copyright e^{-(\lambda+n\mu)t} n\mu, \end{aligned} \tag{4}$$

where © stands for the convolution symbol. Taking Laplace transform on both sides of equation (4), we get

$$\begin{aligned} q^{*}(s) &= \int_{0}^{\infty} e^{-st} q(t)\, dt = \frac{\lambda}{s+\lambda} \times \frac{\mu}{s+\lambda+\mu} + \frac{\lambda}{s+\lambda} \\ &\quad \times \sum_{n=2}^{\infty} \frac{n\lambda^{n-1}\mu}{(s+\lambda+\mu)\ldots(s+\lambda+n\mu)} \\ &= \sum_{n=1}^{\infty} \frac{n\lambda^{n}\mu}{(s+\lambda)(s+\lambda+\mu)\ldots(s+\lambda+n\mu)}. \end{aligned} \tag{5}$$

Using the identity $\dfrac{1}{x(x+b)\ldots(x+nb)} \equiv \dfrac{1}{n!b^{n}} \sum_{j=0}^{n} (-1)' \begin{pmatrix} n \\ j \end{pmatrix} \dfrac{1}{(x+jb)}$, the equation (5) yields

$$q^{*}+(s) = \lambda \sum_{n=1}^{\infty} \frac{1}{(n-1)!}^{(n-1)} \left\{ \sum_{j=0}^{n} \begin{pmatrix} n \\ j \end{pmatrix} (-1)' \frac{1}{(s+\lambda+j\mu)} \right\}. \tag{6}$$

Inverting 6, we obtain explicitly the p.d.f. of ST given by

$$\begin{aligned} q(t) &= \lambda \sum_{n=1}^{\infty} \frac{1}{(n-1)!} \left(\frac{\lambda}{\mu}\right)^{(n-1)} \left\{ \sum_{j=0}^{n} \begin{pmatrix} n \\ j \end{pmatrix} (-1)^{j} e^{-(\lambda+j\mu)t} \right\} \\ &= \lambda \sum_{n=1}^{\infty} \frac{1}{(n-1)!} \left(\frac{\lambda}{\mu}\right)^{(n-1)} e^{-\lambda t} \left\{ \sum_{j=0}^{n} \begin{pmatrix} n \\ j \end{pmatrix} (-1)^{j} e^{-jt\mu} \right\} \\ &= \sum_{n=1}^{\infty} \frac{1}{(n-1)!} \begin{pmatrix} \lambda \\ \mu \end{pmatrix}^{(n-1)} e^{-\lambda t} \left(1-e^{-\mu t}\right)^{n} \\ &= \lambda e^{-\lambda t} \left(1-e^{-\mu t}\right) \sum_{n=1}^{\infty} \frac{1}{(n-1)!} \left(\frac{\lambda}{\mu}\right)^{(n-1)} \left(1-e^{-\mu t}\right)^{n-1} \\ &= \lambda e^{-\lambda t} \left(1-e^{-\mu t}\right) \sum_{n=1}^{\infty} \frac{1}{n!} \left(\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right) \\ &= \lambda e^{-\lambda t} \left(1-e^{-\mu t}\right) \exp\left(\frac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right). \end{aligned} \tag{7}$$

The mode of the frequency curve corresponding to $q(t)$ can be obtained by solving the equation $q'(t) = 0$ or

$$\lambda e^{-\lambda t}\left(1-e^{-\mu t}\right)\exp\left(\tfrac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right)\tfrac{\lambda}{\mu}\left(\mu e^{-\mu t}\right)$$
$$+\lambda\left(-\lambda e^{-\lambda t}\right)\left(1-e^{-\mu t}\right)\exp\left(\tfrac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right)$$
$$+\lambda e^{-\lambda t}\left(\mu e^{-\lambda t}\right)\exp\left(\tfrac{\lambda}{\mu}\left(1-e^{-\mu t}\right)\right)=0.$$

Or $\lambda^2\left(1-e^{-\mu t}\right)\left(e^{-\mu t}\right)-\lambda^2\left(1-^{-\mu t}\right)+\lambda\mu e^{-\mu t}=0.$

Or $\lambda e^{-2\mu t}-\left(2\lambda+\mu\right)e^{-\mu t}+\lambda=0.$

Or $e^{-\mu t}=\dfrac{2\lambda+\mu+\mu\pm\sqrt{\left(2\lambda+\mu\right)^2-4\lambda^2}}{2\lambda}=\dfrac{2\lambda+\mu\pm\sqrt{\mu^2 4\lambda\mu}}{2\lambda}.$

Since $t>0,\ 0<e^{-\mu t}<1$ and so the only possibility is $e^{-\mu t}=\dfrac{2\lambda+\mu-\sqrt{\mu^2+4\lambda\mu}}{2\lambda}.$

Consequently, we obtain

$$t_{\mathrm{mod}\,e}=\frac{1}{\mu}\log\left(\frac{2\lambda}{2\lambda+\mu-\sqrt{\mu^2+4\lambda\mu}}\right). \tag{8}$$

The distribution function $Q(t)$ is given by

$$Q(t)=\int_0^t q(u)\,du=\int_0^t\lambda e^{-\lambda u}\left(1-e^{-\mu u}\right)\exp\left(\frac{\lambda}{\mu}\left(1-e^{-\mu u}\right)\right)du.$$

Putting $v=1-e^{-\mu u},\ dv=\mu e^{-\mu u}du,$

$$Q(t)=\frac{\lambda}{\mu}\int^{1-e^{-\mu t}}v\left(1-v\right)^{\frac{\lambda}{\mu}-1}e^{\frac{\lambda}{\mu}v}\,dv. \tag{9}$$

If $\lambda=\mu,$ then $q(t)=\lambda e^{-\lambda t}\left(1-e^{-\lambda t}\right)\exp\left(1-e^{-\lambda t}\right)$ and

$$\begin{aligned}Q(t) &= \int_0^{1-e^{-\lambda t}}ve^v dv=\left[ve^v-e^v\right]_0^{1-e^{-\lambda t}}=e^{1-e^{-\lambda t}}\left(-e^{-\lambda t}\right)-(-1)\\ &= 1-e^{-\lambda t}e^{1-e^{-\lambda t}}.\end{aligned} \tag{10}$$

In this case, the hazard function $\lambda(t)$ is given by

$$\lambda(t) = \frac{q(t)}{1-Q(t)} = \frac{\lambda e^{-\lambda t}\left(1-e^{-\lambda t}\right)\exp\left(1-e^{-\lambda t}\right)}{e^{-\lambda t}e^{1-e^{-\lambda t}}} = \lambda\left(1-e^{-\lambda t}\right). \tag{11}$$

It can be observed that the hazard rate is increasing monotonically, which agrees with Brookmeyer and Gail (1994). In the next section, the moments of ST are obtained using equation (6).

2.2 The moments of ST

The *k*-th moment of ST is given by

$$E\left[ST^k\right] = (-1)^k\left[\frac{d^k}{ds^k}\left\{q^*(s)\right\}\right]_{s=0}. \tag{12}$$

Consequently, from equation (6) we obtain

$$E\left[ST^k\right] = k!e\frac{\lambda}{\mu}\sum_{j=0}^{\infty}\frac{(-1)^j}{j!\left(\lambda+j\mu\right)^k}\left(\frac{\lambda}{\mu}\right)^j. \tag{13}$$

For the particular case $\lambda = \mu = \lambda$, the mean and variance of ST obtained from equation (13) are given by

$$E[ST] = \frac{e-1}{\lambda} \tag{14}$$

$$Var[ST] = \frac{2e}{\lambda^2}\sum_{n=0}^{\infty}\frac{(-1)^j}{(j+1)(j+1)!} - \left(\frac{e-1}{\lambda}\right)^2. \tag{15}$$

The parameters of $q(t)$ are estimated in the next section by using the method of maximum likelihood.

2.3 Estimation of the parameters of $q(t)$

The likelihood function $L(\lambda,\ \mu)$ for a sample of size n is given by

$$L(\lambda,\ \mu) = \lambda^n \exp\left(-\lambda\sum_{i=0}^{n}t_i\right)\prod_{j=1}^{n}\left(1-e^{-\mu t_j}\right)\exp\left\{\frac{\lambda}{\mu}\left(n-\sum_{k=1}^{n}e^{-\mu t_k}\right)\right\}. \tag{16}$$

The logarithm of L is given by

$$\log_e L = n \log \lambda - \lambda \sum_{i=1}^{n} t_i + \sum_{j=1}^{n} \log \left(1 - e^{-\mu t_j}\right) + \frac{\lambda}{\mu} \left(n - \sum_{k=1}^{n} e^{-\mu t_k} \right). \tag{17}$$

When $\log_e L$ reaches its maximum value, the values of λ and μ satisfy the following simultaneous equations:

$$n(\lambda + \mu) - \lambda\mu \sum_{i=1}^{n} t_i - \lambda \sum_{k=1}^{n} e^{-\mu t_k} = 0 \tag{18}$$

$$\mu^2 \sum_{j=1}^{n} \frac{t_j e^{-\mu t_j}}{1 - e^{-\mu t_j}} + \lambda\mu \sum_{k=1}^{n} t_k e^{-\mu t_k} - n\lambda + \lambda \sum_{k=1}^{n} e^{-\mu t_k} = 0. \tag{19}$$

From equation (18), we obtain

$$\lambda = \frac{n\mu}{\mu \sum\limits_{i=1}^{n} t_j + \sum\limits_{k=1}^{n} e^{-\mu t_k} - n}. \tag{20}$$

Substituting equation (20) into equation (19), we obtain the following transcendental equations for μ:

$$\begin{aligned} \mu \left(\sum_{j=1}^{n} \frac{t_j e^{-\mu t_j}}{1 - e^{-\mu t_j}} \right) \left(\mu \sum_{j=1}^{n} t_j + \sum_{k=1}^{n} e^{-\mu t_k} - n \right) \\ + n \left(\mu \sum_{k=1}^{n} t_k e^{-\mu t_k} - n + \sum_{k=1}^{n} e^{-\mu t_k} \right) = 0. \end{aligned} \tag{21}$$

Equation (21) can be solved using the Newton-Raphson algorithm (Sastry 1994). Accordingly, we put

$$\begin{aligned} \psi(\mu) = \mu \left(\sum_{j=1}^{n} \frac{t_j e^{-\mu t_j}}{1 - e^{-\mu t_j}} \right) \left(\nu \sum_{j=1}^{n} t_j + \sum_{k=1}^{n} e^{-\mu t_k} - n \right) \\ + n \left(\mu \sum_{k=1}^{n} t_k e^{-\mu t_k} - n + \sum_{k=1}^{n} e^{-\mu t_k} \right). \end{aligned} \tag{22}$$

If $\mu^{(0)}$ is an initial approximate value of μ, then the $(l+1)$-th iterate of μ is given by the equation

$$\mu^{(l+1)} = \mu^{(l)} - \frac{\psi\left(\mu^{(1)}\right)}{\psi'\left(\mu^{(1)}\right)},\ l = 0,\ 1, ... \tag{23}$$

3. A stochastic model of the HIV incubation period

Assume that an individual has tested HIV positive for the first time at time $t = 0$. Let the conditional probability that he/she shows the first identifiable symptoms of AIDS during the *n*-th year given that he/she has not shown any symptoms of AIDS in the previous years be given by

$$1 - e^{-n\mu}, \quad n = 1,\ 2, ...,\ \mu > 0 \tag{24}$$

Let IT be the random variable representing the incubation period. In the next section, a one-parameter family of distribution functions of IT is obtained.

3.1 The probability distribution of the incubation period

Let the probability function of IT be defined by

$$p_n = \Pr\{IT = n\}. \tag{25}$$

Then p_n represents the probability that the individual shows the first symptom of AIDS in the *n*-th year. To find p_n, we observe that the individual did not show any symptom in the first year, did not show any symptom in the second year,..., did not show any symptom in the $(n-1)$-th year and shows first symptom in the n-th year. By using the multiplication rule, we obtain

$$p_n = e^{-\mu}e^{-2\mu}...e^{-(n-1)\mu}\left(1 - e^{-n\mu}\right), \quad n = 1, 2, ... \tag{26}$$

Simplifying equation (26) yields

$$p_n = e^{-\frac{(n-1)n}{2}\mu} - e^{-\frac{n(n+1)}{2}}, \quad n = 1,\ 2,\ ... \tag{27}$$

Clearly, we get

$$\begin{aligned}\sum_{n=1}^{n} p_n &= \sum_{n=1}^{n}\left[e^{-\frac{(n-1)n}{2}\mu} - e^{-\frac{n(n+1)}{2}\mu}\right] \\ &= \left(1 - e^{-1\times 2\mu}\right) + \left(e^{-1\times 2\mu} - e^{-2\times 3\mu}\right) + ... = 1.\end{aligned}$$

The mode l of the distribution is given by

$$e^{-\frac{(l-2)(l-1)}{2}\mu}\left(1-e^{-l-1\mu}\right) \leq e^{-\frac{l(l-1)}{2}}\left(1-e^{-l\mu}\right) \leq e^{\frac{-l(l+1)}{2}\mu}\left(1-e^{-(l+1)\mu}\right). \tag{28}$$

The median θ of the distribution is given by

$$1 - e^{-\frac{\theta(\theta+1)}{2}\mu} = \frac{1}{2}. \tag{29}$$

From equation (29), we have

$$\mu\theta\left(\theta+1\right) - 2\log 2 = 0. \tag{30}$$

Solving equation (30), the median is given by

$$\theta = \frac{\sqrt{\mu^2 + 8\log 1} - \mu}{2\mu}. \tag{31}$$

3.2 The moments of incubation period

The mean of IT is given by

$$\begin{aligned}E\left[IT\right] &= \sum_{n=1}^{\infty} np_n \\ &= \sum_{n=1}^{\infty} n\left\{e^{-\frac{(n-1)n}{2}\mu} - e^{-\frac{n(n+1)}{2}\mu}\right\} \\ &= \sum_{n=0}^{\infty} n\left\{e^{-\frac{n(n+1)}{2}\mu}\right\}.\end{aligned} \tag{32}$$

The second moment of IT is given by

$$\begin{aligned} E\left[IT^2\right] &= \sum_{n=1}^{\infty} n^2 p_n \\ &= \sum_{n=1}^{\infty} n^2 \left\{ e^{-\frac{(n-1)n}{2}\mu} - e^{-\frac{n(n+1)}{2}\mu} \right\} \\ &= \sum_{n=0}^{\infty} n \left\{ e^{-\frac{n(n+1)}{2}\mu} \right\}. \end{aligned} \tag{33}$$

3.3 Estimation of the parameter of p_n

Equation (27) represents a one-parameter family of probability distributions and for estimation of the parameter, either the method of moments or the method of maximum likelihood can be used.

3.3.1 The method of moments

Let $t_1,\ t_2, ..., t_m$ be a random sample of size n drawn from a population of incubation times of HIV infected individuals. Then the sample mean is given by

$$\bar{t} = \frac{1}{n} \sum_{n=1}^{m}. \tag{34}$$

Replacing $E\left[T\right]$ by $\bar{t}$ in (32), we have

$$\bar{t} = \sum_{n=0}^{\infty} e^{-\frac{n(n+1)}{2}\mu}. \tag{35}$$

As the incubation time of an HIV-infected individual can never be greater than 100 years, equation (35) can be truncated in the following manner:

$$\bar{t} = \sum_{n=0}^{100} e^{-\frac{n(n+1)}{2}\mu}. \tag{36}$$

An approximate value $\overline{\mu}$ of μ can be obtained from equation (35) by using the Newton-Raphson algorithm.

3.3.2 The method of maximum likelihood

The likelihood function $L(\mu)$ for a sample $\{n_1, n_2, ..., n_m\}$ of size m is given by

$$L(\mu) = \prod_{j=1}^{m} e^{-\frac{(n_j-1)n_j}{2}\mu} \left(1 - e^{-\mu n_j}\right). \tag{37}$$

The logarithm of $L(\mu)$ is given by

$$\log_e L(\mu) = -\frac{\mu}{2} \sum_{i=1}^{m} (n_i - 1) n_i + \sum_{j=1}^{m} \log\left(1 - e^{-\mu n_j}\right). \tag{38}$$

When $\log_e L(\mu)$ reaches its maximum value, the value of μ satisfies the following equation:

$$\frac{\partial L}{\partial \mu} = 0. \tag{39}$$

From equation (39), we obtain

$$\sum_{i=1}^{m} n_i \frac{e^{n_i \mu}}{1 - e^{-n,\mu}} \frac{1}{2} \sum_{i=1} m\left(n_i^2 - n_i\right). \tag{40}$$

By applying the Newton-Raphson algorithm to equation (40), an approximate value $\overline{\mu}$ for μ can be obtained.

3.3.3 The method of median

The value of μ can be estimated from equation (30) for a sample of incubation times, we obtain the sample median θ^* and then replacing θ in equation (30) by θ^*, we have the following equation for a crude estimation μ^* of μ:

$$\mu^* = \frac{2 \log 2}{\theta^* (\theta^* + 1)}. \tag{41}$$

A numerical example to compare the three methods is provided in the next section.

3.3.4 A numerical example

The data of 84 homosexuals and bisexual men analysed in Lui et al. (1988) is used to obtain the incubation periods of twenty one individuals who developed

AIDS prior to the year 1988 (Table 1). Estimates for the value of μ by the three methods are obtained and corresponding expected values and standard deviations are determined. The estimates are then used to test the goodness of fit of the distribution obtained.

Table 1. HIV incidence data of 84 homosexuals

Year of HIV infection	Year of diagnosis									
	1979	1980	1981	1982	1983	1984	1985	1986	Censored	Total
1978	0	0	0	1	0	1	1	0	3	6
1979		0	0	0	0	0	0	1	7	8
1980			0	0	0	1	1	1	9	12
1981				0	2	2	1	5	19	29
1982					1	0	3	0	19	23
1983						0	0	0	2	2
1984							0	0	4	4

From this table, the following incubation times (in years) of 21 persons were obtained as:

4, 6, 7, 7,4, 5, 6, 2, 2, 3, 3, 4, 5, 5, 5, 5, 5, 1, 3, 3, 3.

The sample mean is 4.19 years and the sample median is 4 years. By using the Newton-Raphson algorithm in equation (36), with Table 2, we have the optimal value $\widehat{\mu} = 0.09$ so that the expected value of IT is 4.19 years with a standard deviation of 2.15 years. On the other hand, for the same data of 21 persons, by adopting the Newton-Raphson algorithm in equation (40), we get $\widetilde{\mu}$ so that the expected value of IT is 1.41 years with a standard deviation of 0.59 year. Also, using equation (41), we get $\mu^* = 0.07$ so that the expected value of IT is 4.80 years with a standard deviation of 2.48 years. The three computed values of the parameter μ are listed in Table 2.

Table 2. Values of the parameters of μ

Method	μ	Mean	Standard deviation
Moments	$\widehat{\mu} = 0.09$	4.19	2.15
Maximum likelihood	$\widehat{\mu} = 1.01$	1.41	0.59
Median	$\mu^* = 0.07$	4.80	2.48

Further, by applying the χ^2 test, it was observed that the value of μ obtained by the method of moments fits closely to the observed data. Hence in what follows, we assume $\nu = 0.09$ and proceed to project AIDS incidence by the back-calculation method with a sample data (Bacchetti 1990).

4. The back-calculation and the infection rate

One of the methods used in estimating and projecting the infection rate from AIDS incidence data is the back-calculation method (Brookmeyer & Gail, 1994). It is an important method of constructing rates of HIV infection and estimating current prevalence of HIV infection and future incidence of AIDS (Bacchetti et al., 1993). This method has been used by many mathematical scientists to obtain and predict the AIDS incidence of different populations. Amongst the work done are those of Verdecchia and Mariotto (1995) who modelled past HIV infections in Italy considering the interaction between age and calendar time. Anbupalam et al. (2002) also used the back-calculation method to project future AIDS cases in Tamil Nadu by assuming that the incubation distribution was Weibull and Log-logistic. Ong and Soo (2006) estimated the HIV infection rates and projection in Malaysia while Lopman and Gregson (2008) used the back-calculation method to reconstruct the historical trends in HIV incidence in Harare, Zimbabwe by using mortality data. They also attempted to determine the amount of peakedness of HIV incidence and when the peakedness occurred in Harare, Zimbabwe.

The method in continuous time is based on the convolution equation

$$A(t) = \int_0^t g(s) F(t-s)\, ds \tag{42}$$

where $A(t)$ represents the expected cumulative number of AIDS cases diagnosed by calendar time t, $g(s)$ is the infection-rate at calendar time s and $F(t)$ is the distribution of the incubation period. Equation (42) is a Volterra integral equation for $g(s)$ and has been obtained by noting that an individual can be diagnosed to have AIDS before calendar time t, provided he/she has been infected at some time $s < t$ and has an incubation period less than $t - s$. For a given AIDS incidence data, $A(t)$ can be fitted and a model used for $F(t)$ in 42 so that the rate $g(s)$ can be computed by de-convolving equation (42). Taking Laplace transform on both sides of (42), we have

$$A^*(u) = \frac{g^*(u) f^*(u)}{u} \tag{43}$$

so that

$$g^*(u) = \frac{u A^*(u)}{f^*(u)}. \tag{44}$$

By inverting (44), we obtain the infection rate $g(s)$.

On the other hand, the back-calculation in discrete time is based on the equation

$$E(Y_j) = \sum_{i=1}^{j} g_i p_{j-i+1} \tag{45}$$

where Y_j is the number of AIDS cases diagnosed in the *j*-th year $[j-1,\ j]$, g_j is the number infected in the beginning of the *j*-th year and p_j is the probability that a person who is infected at the beginning of the 1st year is diagnosed with AIDS in the *j*-th year. If A_n denotes the expected cumulative number of AIDS cases diagnosed up to the end of the *n*-th year, then using equation (45), we have

$$A_n = \sum_{j=1}^{n} E\left(Y_j\right) = \sum_{j=1}^{n} \sum_{i=1}^{j} g_i p_{j-i+1}. \tag{46}$$

Equation (46) is analogous to equation (42).

We proceed to illustrate the back-calculation in discrete time with the data used in Bacchetti (1990) where the monthly infection rate and monthly AIDS incidence among gay men in San Francisco in the cohort born from October 1929 through September 1959 were estimated. Taking $t = 0$ to correspond to January 1978 and the time unit as year, the data is given in Table 3 below.

Table 3. Data on AIDS incidence among gay men in San Francisco

j	1	2	3	4	5	6	7	8	9	10	11
Y_j	0	0	1	26	93	278	560	840	1264	1464	1455

Table 4. Probability distribution of the incubation time

n	1	2	3	4	5	6	7	8	9	10
p_n	0.09	0.15	0.18	0.18	0.15	0.11	0.07	0.04	0.02	0.01

For $\mu = 0.09$, the probability distribution of the incubation time is given in Table 4. Following Brookmeyer and Gail (1994), we proceed to obtain the discrete time infection curve. We assume for simplicity that infections occurring in a calendar year are counted at a single time point, for example, January 1 of the year and

$$g\left(2n-1\right) = g\left(2n\right) = \beta_n, \;\; n = 1, 2, ... \tag{47}$$

Equation (47) provides a simple smoothness assumption on the annual infection rate. Consequently, equation (45) leads to the following matrix

equation:

$$\begin{pmatrix} E(Y_1) \\ E(Y_2) \\ E(Y_3) \\ E(Y_4) \\ E(Y_5) \\ E(Y_6) \\ E(Y_7) \\ E(Y_8) \\ E(Y_9) \\ E(Y_{10}) \\ E(Y_{11}) \end{pmatrix} = \begin{pmatrix} 0.09 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.24 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 9.34 & 0.09 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.36 & 0.24 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.32 & 0.34 & 0.09 & 0.00 & 0.00 & 0.00 \\ 0.25 & 0.36 & 0.24 & 0.00 & 0.00 & 0.00 \\ 0.18 & 0.32 & 0.34 & 0.09 & 0.00 & 0.00 \\ 0.11 & 0.25 & 0.36 & 0.24 & 0.00 & 0.00 \\ 0.06 & 0.18 & 0.32 & 0.34 & 0.09 & 0.00 \\ 0.03 & 0.11 & 0.25 & 0.36 & 0.24 & 0.00 \\ 0.01 & 0.06 & 0.18 & 0.32 & 0.34 & 0.09 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} \tag{48}$$

Using the Poisson Regression Analysis (PRA) (Koch et al., 1986; McCullagh & Nelder, 1989), the values of β_j for $j = 1, 2, ..., 6$ are estimated. The method is based on the assumption that the random variable Y_j has a Poisson distribution. Setting $\mu_j = E(Y_j)$, the likelihood function corresponding to the sample $\{n_1, n_2, ..., n_{11}\}$ of $\{Y_1, Y_2, ..., Y_{11}\}$ is given by

$$\varphi(\mu_1, \mu_2, ..., \mu_{11}) = \prod_{i=1}^{11} e^{\mu_i} \frac{\mu_i^{N_i}}{n_i!}. \tag{49}$$

But from equation (48), we have

$$\mu_1 = 0.09\beta_1$$
$$\mu_2 = 0.24\beta_1$$
$$\mu_3 = 0.34\beta_1 + 0.09\beta_2$$
$$\mu_4 = 0.36\beta_1 + 0.24\beta_2$$
$$\mu_5 = 0.32\beta_1 + 0.34\beta_2 + 0.09\beta_3$$
$$\mu_6 = 0.25\beta_1 + 0.36\beta_2 + 0.24\beta_3$$
$$\mu_7 = 0.18\beta_1 + 0.32\beta_2 + 0.34\beta_3 + 0.09\beta_4$$
$$\mu_8 = 0.11\beta_1 + 0.25\beta_2 + 0.36\beta_3 + 0.24\beta_4$$
$$\mu_9 = 0.06\beta_1 + 0.18\beta_2 + 0.32\beta_3 + 0.34\beta_4 + 0.09\beta_5$$

$$\mu_{10} = 0.03\beta_1 + 0.11\beta_2 + 0.25\beta_3 + 0.36\beta_4 + 0.24\beta_5$$
$$\mu_{11} = 0.01\beta_1 + 0.06\beta_2 + 0.18\beta_3 + 0.32\beta_4 + 0.34\beta_5 + 0.09\beta_6$$

and hence substituting of these equations in 49, we find that φ becomes a function of $\beta_1,\ \beta_2, ...,\ \beta_6$. Differentiating $\log_e \varphi$ with respect to β_j and equating the results to 0, the following system of equations is obtained:

$$\sum_{i=1}^{11} \left(\frac{\mu_i - n_i}{\mu_i} \right) \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j = 1, 2, 3, 4, 5, 6. \tag{50}$$

The system of equations 50 does not yield an explicit solution and so an iterative method is used to obtain an approximate solution for $(\beta_1,\ \beta_2,\ ...,\ \beta_6)$ as given below:

$$\widehat{\beta}_1 = 6,\ \widehat{\beta}_2 = 33,\ \widehat{\beta}_3 = 1041,\ \widehat{\beta}_4 = 2583,\ \widehat{\beta}_5 = 3416,\ \widehat{\beta}_6 = 5172.$$

The above values can be used to forecast AIDS incidence on short term. For example, the predicted AIDS incidence in the 12th year is obtained as 6523 by using the following extended equation

$$\widehat{Y}_{12} = \widehat{\beta}_1 (p_{11} + p_{10}) + \widehat{\beta}_2 (p_9 + p_8) + ... + \widehat{\beta}_6 (p_{3\_} p_2)\,. \tag{51}$$

Acknowledgements

The authors wish to thank the referees for their valuable comments. Thanks are also due to NRF for funding this project.

References

ANBUPALAM, T., RAVANAN, R. & VENKATESAN, P. (2002). Back-calculation of HIV and AIDS in Tamil Nadu. *Biostatistical aspect of Health and Epidemiology*, pp. 232–243.

ANDERSON, R.M & MAY, R.M. (1991). *Infectious diseases of humans: dynamics and control*. Oxford and New York: Oxford University Press.

BACCHETTI, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *Journal of the American Statistical Association*, Vol. **85**, No. 412, pp. 1002–1008.

BACCHETTI, P., SEGAL, M.R., & JEWELL, N.P. (1993). Back-calculation of HIV infection rates. *Statistical Sciences*, Vol. **8**, No. 2, pp. 82–119.

BROOKMEYER, R & GAIL, M.H. (1994). *AIDS epidemiology: A quantitative approach*. Oxford University Press, US.

CHEVRET, S., COSTAGLIOLA, D., LEFRERE, J.J. & VALLERON, A.J. (1992). A new approach to estimating AIDS incubation times: results in homosexual infected men. *Journal of Epidemiology Community Health*, Vol. *46(6)*, pp. 582–586.

KOCH, G.C., ATKINSON, S.S. & STOKES, M.E. (1986). *Poisson regression. In Encyclopedia of Statistical Sciences*, edited by Samuel Kotz and Norman L. Johnson. Wiley New York, Vol. 7.

LEE, S. (1999). Estimation of the maturity of HIV and the incubation period of AIDS patients. http://www.tilastokeskus.fi/isi99/proceedings/arkisto/varasto/lee_0375.pdf

LOPMAN, B. & GREGSON, S. (2008). When did HIV incidence peak in Harare, Zimbabwe? Back-calculation from mortality statistics. *PLos ONE online journal*, Vol. **3(3)**: e1711 (http://www.plosone.org).

LUI, K.J, DARROW, W.W & RUTHERFORD, G.W. (1988). A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science*, Vol. **240**, No. 4857, pp. 1333–1335.

McCULLAGH, P. & NELDER, J.A. (1989). *Generalized linear models*. Chapman and Hall. 2nd edition.

MEDLEY, G.F., BILLARD, L., COX, D.R. & ANDERSON, R.M. (1988). The distribution of the incubation period for the acquired immunodeficiency syndrome (AIDS). *Proceedings of the Royal Society of London, Series B, Biological Sciences*, Vol. **233**, No. 1272, pp. 367–377.

ONG, H.C. & SOO, K.L. (2006). Back-calculation of HIV infection rates in Malaysia. *The Medical Journal of Malaysia*, Vol. **61(5)**, pp. 616–620.

RAO, A.S.R.S. & KAKEHASHI, M. (2005). Incubation – Time distribution in back-calculation applied to HIV/AIDS data in India. *Mathematical Biosciences in Engineering*, Vol **2(2)**, pp. 263–277.

SASTRY, S.S. (1994). *Introductory methods of numerical analysis*. Prentice Hall India, 3rd edition.

VERDECCHIA, A. & MARIOTTO, A.B. (1995). A back-calculation method to estimate the age and period HIV infection intensity, considering the susceptible population. *Statistical Medicine*, Vol. **14(14)**, pp. 1513–1530.

Manuscript received, 2010.02.02, revised, 2010.07.20, accepted, 2010.12.02.