CHAPTER V
DISCUSSION

Mathematical and statistical models can be used to estimate HIV incidence with reasonable levels of confidence. Several models have been developed over time that generally depend on reliable HIV prevalence data and on assumptions about survival after infection. These methods include, for example, dynamical models, demographic models, back-calculation techniques and birth cohort methods (Gregson et al. 1996; Williams et al. 2001; Downs et al. 2000). Among the most commonly used methods to derive HIV incidence is the UNAIDS/WHO recommended Estimation and Projection Package (EPP) and Spectrum AIDS Impact Model. EPP fits an epidemiological model to observed HIV prevalence data collected over time using maximum likelihood procedures, while Bayesian techniques are employed to estimate the level of uncertainty around the epidemic curve (Brown et al. 2008). EPP version 2009 calculates incidence from HIV prevalence by taking account of the number of people receiving antiretroviral therapy. Together with the epidemic curve produced in EPP, the Spectrum software uses demographic data, information on adult and child treatment coverage and assumptions about the epidemiology of HIV to generate estimates of national (adult and child) HIV prevalence, incidence, mortality and treatment needs (Stover et al. 2008). These methods have been used in more than 120 countries worldwide to provide national, regional and global estimates of the impact of HIV.
A model developed by the South African Actuarial Society (ASSA) has been used by some countries in southern Africa (South Africa and Botswana) to project the demographic impact of HIV, with results consistent with those obtained from Spectrum (Dorrington et al. 2002).

Hallett et al., on behalf of the ALPHA Network, have recently developed methods to estimate HIV incidence by age in the general population using successive rounds of cross-sectional prevalence data (Hallett et al. 2008). The methods examine the change in HIV prevalence in a cohort observed at two time points, allowing for changes due to new infections and mortality among infected and uninfected persons, using either cohort mortality rates or using the distribution of survival after HIV infection. The modelled estimates of incidence showed good agreement with those obtained directly from three community-based cohorts in Africa (Hallett et al. 2008). The model, which has been extended to take account of the effect of antiretroviral therapy on prevalence, has already been applied to data from several countries and is highly recommended for use in other countries that have done (or are planning to do) more than one national HIV survey (Hallett et al. 2010). Another method that has recently been used widely is the UNAIDS model to estimate the distribution of new HIV infections by modes of transmission for a one-year period (Gouws et al. 2006). The modes of transmission model was developed to guide prevention activities in countries by identifying those groups at highest risk of HIV infection and to make countries aware of changing patterns of HIV risk. The model requires data on risk behaviour, population sizes, prevalence of HIV and sexually transmitted
infections by risk group, and application is therefore limited to those countries
that have the required data. In the absence of reliable data, uncertainty ranges are
likely to be very wide.

The Asian Epidemic Model (AEM) was developed to assess infection patterns in
risk groups over time (Brown et al. 2008). The AEM model replicates the key
processes driving HIV transmission in Asia and offers opportunities to explore
the effectiveness of different intervention and care programmes. The model
requires extensive behavioural input data over time and the application has been
limited to only a few countries for which such data are available.

Back-calculation techniques have been used mainly in developed countries with
reliable data on the number of AIDS diagnoses over time and with information on
the distribution of the incubation period. More recently, extended back-
calculation methods have been developed to overcome some of the shortcomings
of the original methods, such as the modification of parameters by the use of
antiretroviral therapy, by utilizing more information about AIDS cases than
before (Campsmith et al. 2010; Hall et al. 2008). Our study is the first research
done for finding the best fitted statistical distribution for incubation time of HIV
and Back-calculation for estimating the AIDS cases in Kerala.

19,905 persons were tested positive for HIV infection until now in Kerala. The
highest number of positive cases was reported in Thiruvananthapuram with total
4,034 cases: 2,126 men, 1,489 women and 219 children. Thrissur came second
with 3,426 cases: 1,865 men, 1,376 women and 185 children, while Kozhikode
reported 2,953 cases: 1,694 men, 1,097 women and 162 children. A total of 899 children were registered at the Anti-Retroviral Therapy (ART) centres, the KSACS officials said in reply to the RTI query filed by advocate D B Binu. According to the data of 2011, 61 children were found to be positive, with 12 each in Thiruvananthapuram and Thrissur and 10 in Kozhikode. The rate of infection is coming down in the state even though more cases are being detected because of the awareness among the public (The Times Of India 2012). Using the back calculation method, in our study we projected cumulative number of AIDS cases of Kerala up to 2010. According to back calculation, there were 62039 AIDS cases in 2007 and 43606 in 2010. It is slightly higher than the numbers revealed by KSACS(KSACS 2007). It might because of the software programme EPP used by the KSACS, which will be using the parameters of incubation time reported by non Keralite study. We applied Weibull model, Gamma model, Log normal model, logistic model and Fisher-Tippett model in AIDS incubation time. In these five models considering the weibull family and gamma family only, we fitted three distributions for each family. They are weibull 1, weibull 2, weibull 3, gamma 1, gamma 2 and gamma 3. Weibull 2 distribution was selected for best fitted among all the other distributions for total, male and female incubation time because of its distribution function nature was suitable to the nature of AIDS incubation time and statistical significance(p<0.01). There were several other studies done in the other part of world for finding the incubation time of AIDS. Weibull and gamma models were the most commonly used for back calculation approach. Between the two, Weibull model is a popular candidate for HIV incubation period because of its nice properties viz., it is proportional hazard as
well as accelerated failure time model. The earliest studies of Weibull incubation period have been attempted (Lui et al. 1988; G. F. Medley et al. 1987) and fitted Weibull model to study the incubation period distribution for transfusion associated AID cases. These parameter values correspond to a median incubation period of 4.3 years. The incubation period of patients infected by blood transfusion was studied (Medley et al. 1987) using Weibull distribution for Children (0-4 years) and estimated the parameters as follows: \( \alpha = 1.9390, \lambda^\alpha = 0.1566 \), for others (5-59 years) as \( \alpha = 2.3960, \lambda^\alpha = 0.0048 \). Boldson et al. (Venkatesan 2006) used gamma, Weibull and lognormal models for incubation time of cohort study from San Francisco AIDS cases. The fitted Weibull model for their data is given by \( F(t) = 1 - \exp(-0.001296 t^{2.5}) \). The Weibull HIV incubation model used by Anderson et al. (1986) is given by \( F(t) = 1 - \exp(-0.1190 t^{1.9974}) \). Brookmeyer and Goedert used the Weibull incubation period distributions based on the study of hemophiliacs over 20 years of age (Brookmeyer & Goedert 1989). The fitted Weibull model for their data is given by \( F(t) = 1 - e^{-0.0021 t^{2.516}} \). This estimate corresponds to a median incubation of 10 years. Based on 732 HIV-positive hemophiliacs enrolled in Italian registry, Chiarotti et al. estimated the incubation distributions assuming three different parametric models: uniform, uniform in three sub intervals and truncated Weibull under two approaches namely the median and median of three random values (Chiarotti et al. 1994). There are altogether six different approaches to estimate the incubation time of individuals. They found that the incubation times obtained using first and third are similar. Therefore they reported only four estimates. The estimates of the four models parameters are \( \alpha = 2.9, 2.4, 2.9, 2.6 \) and \( \lambda^\alpha = 0.003668, 0.001039, 0.000446, 0.000845 \). The median
incubation times are 13.5, 15, 12.6 and 13.3 respectively. Munoz and Xu using Multicenter AIDS Cohort Study (MACS) estimated this Weibull model (Muñoz & J. Xu 1996) $F(t) = 1 - e^{-0.052087 t^{1.285347}}$. The median incubation period using this model is 7.5 years. Other important studies also used the Weibull model for HIV incubation period (Mode et al. 1988; Rosenberg & Gail 1990). Gamma distribution is the other important parametric distribution used to model incubation period of HIV/AIDS. One of the earliest studies used gamma model for incubation period of HIV estimated parameter for the gamma models are as follows (Medley et al. 1987). $K=2.669, 2.473$ and $\alpha=0.911$ and 11.001. The parameters estimated from gamma model (Venkatesan 2006) based on the San Francisco AIDS data gave $k = 3.130$ and $\sigma = 5.715$ years. Freund and Book (Freund & Book 1990) fitted gamma model with $k = 3$ (Erlang form) to the San Francisco AIDS data and the estimate obtained for the parameter $\sigma = 2.660$ years. Lawless and Sun used the log-logistic model for HIV incubation period, in addition Chiarotti et al. used log-logistic model and generalized exponential model for their data (Chiarotti et al. 1994; Venkatesan 2006). The log-normal distribution has been used (Rees 1987) for HIV incubation period. Recent studies have shown that lognormal distribution fits well than Weibull model (Muñoz & Xu 1996; Chiarotti et al. 1994). This has been shown that the three parameters generalized log-logistic distribution with fits better than the log-logistic distribution for data on cancer survival analysis (Venkatesan 2006). Stacy introduced a generalization of gamma distribution with three parameters. This model is a generalization of many survival distributions. The convolution of exponential distribution has been used as incubation model for HIV (Singh et al.
Longini et al. used a staged Markov model to estimate the distribution and mean length of the incubation period from a cohort study of 603 HIV infected individuals who have been followed through various stages of infection. They used the generalized gamma model to describe the transition probabilities of the Markov model (Longini et al. 1989).

It is generally observed that the short term projected AIDS cases do not vary much across various infection densities and incubation period distributions. But the minimum size of the epidemic and HIV incidences are highly variable across the infection densities and incubation period distributions. The projected AIDS cases within an infection density across various incubation distributions are found to be very stable. But across the infection densities the variation is observed to be high. The projected AIDS estimates obtained using the weibull incubation time density and back calculation method seems to be more plausible compared to the estimates obtained using the other models. It is to be noted that these estimates are based on the unadjusted AIDS incidence data. These estimates may not be the correct number of AIDS cases that may develop in Kerala during these periods. The exact number of AIDS cases that may develop in Kerala will be certainly higher than these figures and hence these figures can be taken to be a lower bound for possible number of AIDS cases.

A plot is a graphical representation of the collected data (independent and dependent variables) involved in a study. The association between these variables are then assessed by connecting the `points' with a line. Though very true, this association cannot be relied upon to predict the future trend of this data. Now a
`model', which `fits best' to the observed data has to be worked out. This is then `fitted' and used to replace the existing set of data points as `the appropriate model'. After `modelling' the observed data, this model can be used to predict future trend of the dependent variable for a given change in the other. The foregoing statement covertly mentions several requirements which often ensure confident achievement in any subsequent extrapolation from the model. The model selected must be the most appropriate for the collected data. A usable and understandable curve-fitting method is to be available from which the model facts those are reflective of future behaviour can be obtained(Sathian & Sreedharan 2012; Sathian et al. 2010a; Sathian et al. 2010b; Sathian et al. 2011a; Sathian 2011b; Sathian et al. 2011c). Timely and accurate monitoring of the HIV epidemic requires measures of incidence, that is, the number of new infections in a defined population that occur during a defined time period. Unfortunately, longitudinal studies that have traditionally provided incidence measures are costly, time consuming, logistically complex, and may be subjectively biased, differential loss to follow-up, or an intervention effect(Laga et al. 1994; Brookmeyer et al. 1995). As a result, public health agencies have generally relied on surveys that measure HIV prevalence, the proportion of persons at a specified point in time that are infected, to track the epidemic.

Anderson & May 1992b; Mode et al. 1988; Isham 1991; Tan & Xiang 1999; Jewell et al. 1992; Brookmeyer & Gail 1994; Cazelles & Chau 1997; Brookmeyer & Gail 1986). But there are several uncertainties in these methods and the lack of correct information regarding the HIV data also affect the efficacy of these models (Sathian 2011). Sathian and Sreedharan used curve fitting method for forecasting several infectious and non infectious diseases (Sathian & Sreedharan 2012; Sathian et al. 2010a; Sathian et al. 2010b; Sathian, et al. 2011a; Sathian 2011b; Sathian et al. 2011c). Statistical modelling of HIV cases in Nepal showed cubic model best fitted with 99% accuracy (Sathian et al. 2011a). In our present study, Cubic Model was the best fitted model according to the residual, R^2, and p-value. But it had less reliability because of the shape and mathematical equation of the curve. Graph 4.3.1 depicts, Quadratic model provided closely fitted curve for Adult and children HIV cases compared to cubic model. From the Graph 4.3.1 & Table 4.3.2, 2006 and 2007, 2008 and 2009 cases were same but cubic model couldn’t adjust with this trend. So, it was under forecasted more and given the prediction up to 2014. Using the curve fitting method, we estimated the number and trend of Adult and children HIV cases at India from the year 1990 to 2020.

NACO 2010-11 Annual Reports showed that the estimated adult prevalence of HIV were 0.31 percent with 23.9 lakh people infected, of which, 39 percent were female and 3.5 percent were children based on HIV Sentinel Surveillance 2008-09 (NACO 2011). Our study showed a declining trend of adult and children HIV cases and in 2020, it will be 9.2 lakhs. We got 2.27 million adult and children
HIV cases in 2012, which is similar but little higher than the findings in the Rao et al. study to project the HIV cases in India (Rao et al. 2009). Behavioral intervention was the reason for declining HIV trends in Southern parts of India, some countries in Africa, and Thailand (Kumar et al. 2006; Asiimwe-Okiror et al. 1997; Alary et al. 2002). According to NACO, 0.41 percent was HIV prevalence in 2000; 0.36 percent in 2006, and 0.31 percent in 2009. All the high prevalence states show a clear declining trend in adult HIV prevalence. However, the low prevalence states of Chandigarh, Orissa, Kerala, Jharkhand, Uttarakhand, Jammu & Kashmir, Arunachal Pradesh and Meghalaya show rising trends in adult HIV prevalence in the last four years (NACO 2011). Our study hereby establishes the applicability of statistical modelling in predicting the Adult and children HIV cases in the Indian context.

### 5.1 MAJOR CHALLENGES IN THE APPROACHES USED TO ESTIMATE HIV/AIDS NUMBERS

Antenatal clinic surveys, population based surveys, surveys in high risk subpopulations, and Back-calculation are the four main approaches for estimating population prevalence rates of HIV. Major sources of error and problems with antenatal clinic surveys are because of the representativeness, females only, reproductive age, sexually active period, limited catchment area coverage (e.g., weighted toward urban areas), and uncertainties in adjustment factors. Because of the bias related to non-response, population based surveys requires large sample sizes and hence the expense are relatively more. The errors related to surveys in high risk subpopulations include representativeness, non-response bias and
uncertainty in sizes of subpopulations. Finally, Back-calculation method has the errors arose from the incompleteness in AIDS surveillance data, incubation period, treatment effects, imprecision in recent HIV incidence and requires assumptions about patterns of HIV screening (Brookmeyer 2010; Le Vu et al. 2008; Marston et al. 2008).

Cohort studies, serial HIV prevalence and biomarkers in cross-sectional surveys are the approaches for estimating current incidence rates of HIV.

5.1.1 Cohort studies

Cohort studies of uninfected persons have been widely employed to estimate HIV incidence rates. Karon et al. for a summary incidence from various populations in the United States (Karon et al. 2001). While cohort studies are useful for estimating incidence in selected subpopulations, public health agencies have not generally relied on them for obtaining national estimates of HIV incidence, because of the costs and logistical difficulties involved in following representative national samples of adequate size.

There are important sources of error in cohort studies that could bias incidence rates, even in selected subpopulations. One error, selection bias, arises if persons who agree to participate and return for follow-up visits are not representative of the target population. Another possible source of error is that follow-up visits could affect the HIV incidence rate through repeated exposure to counseling (such as the promotion of condom use or safe sex or other prevention messages). This phenomenon has been called the “adherence effect,” because it occurs
among persons who adhere to the scheduled follow-up visits (Brookmeyer 2009). For example, in the ZVITAMBO study, Piwoz et al. reported increasing HIV knowledge with increasing exposure to an HIV education and counseling program in a cohort of new mothers in Zimbabwe (Piwoz et al. 2005).

5.1.2 Serial prevalence surveys

The basic idea of serial prevalence surveys is to infer HIV incidence (annual numbers of new infections in a population) from changes in absolute HIV prevalence (‘absolute HIV prevalence’ refers here to the numbers of persons living with HIV rather than the percentage infected) at 2 or more points in time.

The approach of estimating HIV incidence from serial prevalence surveys may find increased utility, because surveys such as the Demographic and Health Surveys are planned to be repeated in some countries every 5 years. However, an important drawback of this approach is that it requires critical inputs about mortality and migration, which can be very uncertain. Furthermore, nationally representative surveys realistically can only be conducted once every several years, and as such the approach is unlikely to be able to provide timely information about current trends in HIV incidence (Brookmeyer 2010).

UNAIDS uses a variation of this approach to derive HIV incidence by analyzing trends in prevalence rates at antenatal clinics (Ghys et al. 2006). The UNAIDS approach consists of fitting a smooth parametric curve to the time series of prevalence rates using a computer software program, the Estimation and Projection Package (Ghys et al. 2004). HIV incidence rates are calculated using
additional mortality and demographic assumptions implemented in another software program, SPECTRUM (Stover 2004). There are several uncertainties in this approach. First, as was discussed above, prevalence trends among pregnant women attending public antenatal clinics may not be a useful surrogate for the general population.

Second, mortality and migration assumptions are uncertain. Third, the curve-fitting technique used by UNAIDS to smooth the time series of antenatal care HIV prevalence rates uses a particular mathematical model that allows only a single peak (or mode) in the time trend. Accordingly, the mathematical model currently used by UNAIDS to smooth the noisy time series of antenatal care prevalence rates does not have the flexibility to detect re-emerging sharp upturns in HIV incidence. Some Bayesian approaches have been suggested to account for model uncertainty (Alkema et al. 2008). Some related statistical approaches to inferring HIV incidence from trends in HIV prevalence have also been discussed by Williams et al. and Hallett et al. (Williams et al. 2001; Hallett et al. 2008).

5.1.3 Biomarkers in cross-sectional surveys

The basic idea of the biomarker approach to estimating HIV incidence rates is to use a biomarker to identify persons who were recently infected (Brookmeyer & Quinn 1995). This approach requires a cross-sectional survey of a representative sample of persons in the population from whom serum specimens were collected at a single point in time, which is in contrast to the cohort approach, which requires that multiple specimens be collected longitudinally over time. The
biomarker is used to identify persons who are in the ‘‘window’’ period—a period of time shortly after incident infection occurs.

The biomarker approach relies on the epidemiologic relation that the prevalence of a condition is equal to the incidence multiplied by the mean duration of the condition.

Here, the ‘‘condition’’ refers to the window period. Then the incidence rate $I$ is estimated from the equation

$$I = \frac{p}{\mu} \times 100$$

Where $P$ is the proportion of persons in the window period among all persons who are either uninfected or in the window period and $I$ is the mean window period. If the mean is expressed in years (days), then $I$ is expressed as the percentage per year (per day) of the uninfected population that becomes infected.

The duration of the window period is not fixed but rather is random and has a probability distribution. Individuals may have window periods either above or below the mean window period.

The biomarker approach has been used in both developed and developing countries to estimate HIV incidence. However, there have been reports, especially from Africa, that cohort estimates of incidence are lower than biomarker (BED) estimates, raising the concern that the BED approach systematically overestimates HIV incidence (Karita et al. 2007). In fact, UNAIDS issued a cautionary statement about the use of the BED assay to estimate HIV
incidence. These reports raise questions as to why cohort estimates do not agree with the biomarker estimates.

Major sources of errors and problems with cohort studies are representativeness (selection bias), low follow-up rates and adherence effects expense. Representativeness, errors in assumed mortality rates, unaccounted for migration effects, model for prevalence curve (Especially problematic if based on a time series of serial prevalence surveys from sentinel populations, such as pregnant women attending antenatal care clinics), time interval between surveys (Especially problematic if based on large national prevalence surveys with at least several years between surveys) and expense are the reason for the errors in serial HIV prevalence surveys. Representativeness, mean window period (may depend on strain, population), laboratory errors, impact of advanced HIV disease and antiretroviral therapy, window period distribution and assay nonprogressors are the reason of errors for biomarkers in cross-Sectional Surveys(Brookmeyer 2010).

5.2 LIMITATIONS OF THE RESEARCH

Restricted availability of data on HIV/AIDS in Kerala is a serious limitation on this research. The nature of available data, to a great extent, tailored the direction of this thesis. The data were collected on all HIV tested positive cases between 2005 and 2010, thus omitting the first reported cases in 1987 and data up to 2004. Also, there was unavailability of information about the epidemic in the recent years (2010 - 2012). The aggregation of the data at state level made it impossible to conduct more detailed analysis. The study and modelling of the trend of the epidemic across the various demographic strata of the Kerala state was hindered
by the non stratification of the data. It is strongly recommended that data be placed on the public domain after striping all patient identities and made more accessible to researchers. Aggregation of data should be avoided as much as possible; at least, data should be published by sex and age for each of the districts of Kerala.

The back-projection models adopted an incubation period distribution estimated from a small sample study. It is uncertain to what extent this distribution represents the Keralite epidemic scenario. The estimates from the parametric back calculation highly depend on the incubation time parameters. Another limitation of this study is that the various staging models incorporating effect of therapy, different risk groups and other methods for reporting delays and underreporting were not considered due to non-availability and non-accessibility of the necessary data.