Estimation of the Incidence of Disease with the Use of Prevalence Data

Klaus Langohr
Departament d'Estadística i Investigació Operativa
Universitat Politécnica de Catalunya
Pau Gargallo, 5
E-08028 Barcelona

Abstract: The incidence of a disease can be easily estimated from serial prevalence data. However, difficulties usually arise in the case of prevalence data measured only once. This paper presents and assesses various methods for estimating the incidence of a disease in the latter case. The focus is on how applicable the methods are to the estimation of the incidence of the HIV-epidemic in some parts of Africa where data from serial cross-sectional surveys are hard available.

1 Introduction

The prevalence of a disease is its relative frequency in a population. It therefore corresponds to the probability of being affected by the disease. The incidence of a disease is the proportion of new cases of disease within a determined period in the population. This can be distinguished between cumulative incidence and the incidence rate or density. The cumulative incidence is defined as the ratio of the number of affected people and the population size within a time interval. It gives the probability of people falling ill during such a time interval. The incidence rate, on the other hand, is the ratio of the number of people falling ill and the average population at risk. It states the actual risk of catching the disease (Kreienbrock and Schach [1995]; Chap. 2).

Whereas prevalence does not cover cured people and the deceased, incidence considers all cases of the disease. Consequently, for the planning of clinical or epidemiological studies, knowledge of the incidence of the disease of interest is essential, as well as the number of persons required in a vaccine trial. It is well-known that the future development of a disease can be predicted by means of incidence estimation. Also, the effects of measures taken against the disease can be statistically compared (Gregson et al. [1996]; Ades and Nokes [1993]; Podgor et al. [1983]; Heyward et al. [1994]).

The use of prevalence data as an estimator for the incidence is not appropriate in many situations. This is because prevalence is a function of both incidence and duration of disease. For example, the prevalence of a chronic disease might be very high even when its incidence is low. It is important to note that constant prevalence data over time do not usually give an indication of the development of the incidence (Saidel et al. [1996]; Leske et al. [1981]; Freeman and Hutchison [1980]).

Due to the high number of reported AIDS-cases in some parts of Africa and Asia, it is of particular importance therefore to have reliable data about the spreading of the HIV-infection. However, financial and logistic restrictions often do not permit large and well planned cohort studies from which HIV-incidence could be easily estimated. In the case of cohort studies, the proportion of new infections in the given time intervals is an estimator for the cumulative incidence. Instead of expensive cohort studies, low cost cross-sectional studies are generally conducted. Cross-sectional studies are much easier to conduct because they do not require personal information for the identification of the study participants (Ades and Nokes [1993]; Brookmeyer et al. [1995]).

Available methods for estimating the incidence, such as backward calculation by means of AIDS-prevalence data and mathematical models, have several drawbacks. Backward calculation, for instance, cannot give a reliable estimation of the present HIV-incidence, because the data used are not reflected in present AIDS-prevalence data. The mathematical models require detailed information on many parameters which are usually not available (Saidel et al. [1996]; Brundage et al. [1990]; Brookmeyer and Quinn [1995]).

In order to estimate the incidence of a disease by means of prevalence studies in which measurement is done only once, several approaches have been recently developed, especially with respect to the HIV-epidemic. This paper presents an assessment of these new approaches with a focus on their relevance to the AIDS-epidemic.

2 Incidence Estimation by Means of Prevalence Data

In this section several deterministic and stochastic approaches as well as data conditions will be represented. Detailed descriptions of these methods can be found in the literature cited in the text.

The following notation will be used:

N	Population size,
N_k	Population size during age interval k,
P	Prevalence,
P_k	Prevalence during age interval k,
$P(a)$	Prevalence at age a,
$P_t(a, a{+}r)$	Prevalence during age interval [a, a+r] at time t,
P_N	Absolute prevalence number,
I	Cumulative incidence,
I_k	Cumulative incidence during age interval k,
$I(a)$	Cumulative incidence at age a,
ι	Incidence rate,
ι_k	Incidence rate during age interval k,
$\iota(a)$	Incidence rate at age a,
$\iota_t(a)$	Incidence rate at age a and time t,
R	Net probability for developing disease,
R_k	Net probability for developing disease during age interval k,
q_k	Probability of dying during age interval k,
$p^+_{x,a}$	Probability of surviving HIV-related mortality until age a after infection with HIV at age x,
$\pi^+_{a,a+r}$	Probability of surviving the age interval [a, a+r] from HIV-related mortality.

There are two kinds of incidence estimation — the estimation of the cumulative incidence and the estimation of the incidence rate. Cumulative incidence is also termed as the crude probability for falling ill since it does not take in account the probability for dying. The incidence rate, on the other hand, is termed as net risk of catching a disease as it refers only to the healthy population.

If a constant incidence rate can be assumed for a time interval, then the periods until falling ill are distributed exponentially. In case of short time intervals of length Δ and a low incidence rate, cumulative incidence (I) and incidence rate (ι) are interrelated as follows (Kreienbrock and Schach [1995]; pp. 20):

$$I = 1 - e^{-\iota \cdot \Delta} \approx \iota \cdot \Delta. \tag{1}$$

2.1 Interrelation of Prevalence and Incidence

Prevalence, incidence and the duration of a disease in a steady-state population are interrelated in such a way that any two of the three quantities may be used to obtain the third (Freeman and Hutchison [1980]).

Steady-state assumptions are the following:

1. Incidence and the distribution of duration of illness within the population remain constant over time; and

2. The total population of affected and unaffected individuals, respectively, remains constant over time.

Let D be the duration of the disease and E(D) the corresponding expectation value. Then:

$$\mathrm{P_N} = \mathrm{P} \cdot \mathrm{N} = \mathrm{I} \cdot \mathrm{E(D)}, \tag{2}$$

and

$$\mathrm{P} = \frac{\iota \cdot \mathrm{E(D)}}{1 + \iota\, \mathrm{E(D)}} \;\Leftrightarrow\; \frac{\mathrm{P}}{1 - \mathrm{P}} = \iota \cdot \mathrm{E(D)}. \tag{3}$$

It follows that when the steady-state assumptions are fulfilled, cumulative incidence and incidence rate can be calculated from prevalence data if the expectation value of the duration of the disease is given. Given $\hat{P}$, an estimation for P, estimations for I and ι, $\hat{I}$ and $\hat{\iota}$ respectively, can be obtained with the use of equations (2) and (3).

Equation (2) is exact if incidence and the distribution of duration of the disease remain constant over time. The constant condition for both quantities must hold for as long as the longest disease duration included in the distribution. This is because the number of affected people can consist of all new cases within that time period. The incidence rate can then be estimated by means of equation (3). This equation is exact when both steady-state assumptions are met.

An application of the above interrelations of incidence and prevalence with regard to the HIV-epidemic is provided by Brookmeyer and Quinn (1995). In the setting of a cross-sectional study in India, more than 1900 people were tested for HIV. In addition, seronegatives were tested for the p24 antigen. This antigen can be detected in HIV-infected individuals already during the preantibody period. The expectation value of the duration of the p24 antigen period before seroconversion is estimated by the collected data, after which equation (2) is used for the estimation of the cumulative incidence. The calculation of confidence intervals is based on the assumption that the number of p24 antigen prevalent individuals follows a Poisson distribution.

2.2 Incidence Estimation from Age-Specific Prevalence Data

In the following, approaches developed by Leske et al. (1981, 1986) for diseases which do and do not affect mortality risk are presented. The approach of Elandt-Johnson and Johnson (1980) is also presented. It is important to note that all approaches assume disease to be irreversible.

2.2.1 Equal Mortality for Affected and Unaffected Individuals

A deterministic approach is chosen in this case of equal mortality for affected and unaffected individuals. It is assumed that disease incidence and population composition (regarding risk factors for disease) remain constant over time (Leske et al. [1981]).

The starting point of the model is the total number of affected individuals at the beginning of the age interval k+1. This number consists of all individuals who either were already affected at the beginning of the age interval k or developed the disease during the interval and, in both cases, who did not die during the age interval k:

$$\mathrm{N_{k+1}P_{k+1} = N_kP_k(1 - q_k) + I_k(N_k - N_kP_k)(1 - q_k)}. \tag{4}$$

Using $\mathrm{N_{k+1} = N_k(1 - q_k)}$, the cumulative incidence can be calculated as follows:

$$\mathrm{I_k = \frac{P_{k+1} - P_k}{1 - P_k}}. \tag{5}$$

Usually, the knowledge of the incidence rate ι_{k} is of higher interest since ι_{k} refers to the risk of developing the disease in the presence of competing mortality (Ederer [1964]). It is the ratio of the number of new cases and the average risk population in the k-th age interval.

In the following equation, $\mathrm{h_k}$ represents the proportion of new cases who die out of all (theoretical) individuals who develop the disease *and* die during this age interval:

$$\iota_{\mathrm{k}} = \frac{\mathrm{I_k(N_k - N_kP_k)(1 - q_k) + I_k(N_k - N_kP_k)q_kh_k}}{\mathrm{(N_k - N_kP_k)(1 - \frac{1}{2}q_k)}}. \tag{6}$$

The denominator in (6) indicates that periods from the beginning of the age interval until death are supposed to be distributed uniformly. The incidence rate reduces to:

$$\iota_k = \frac{I_k \left[1 - q_k(1 - h_k)\right]}{(1 - \frac{1}{2}q_k)}. \tag{7}$$

The value of ι_k is equal to the cumulative incidence I_k, if $h_k = \frac{1}{2}$, which means that periods until death and until developing the disease follow the same distribution. If disease-free periods and survival periods are assumed to be exponentially distributed with different parameters, then:

$$\iota_k = \frac{\ln(1 - I_k)}{\ln(1 - q_k) + \ln(1 - I_k)} \frac{1 - (1 - I_k)(1 - q_k)}{1 - \frac{1}{2}q_k}. \tag{8}$$

Using equation (5), in this case the incidence rate is calculated by means of the prevalence and the probability of dying. The variance of the corresponding estimator is derived by the delta method (Leske et al. [1981]).

Leske et al. (1981) and Podgor et al. (1983) have used the above equations to estimate cumulative incidence and incidence rate in connection with the investigation of several eye diseases in England and in the United States, respectively. In both studies there is hardly any difference between $\hat{I}$ and $\hat{\iota}$ for the chosen age intervals, which is supposed to be the case due to the low probability of dying.

2.2.2 Different Mortality for Affected and Unaffected Individuals

Similar to section 2.2.1 the starting point for the case of different mortality for affected and unaffected individuals (q'_k and q_k, respectively) is equation (4). In this case, however, the total number of unaffected individuals at the beginning of age interval k+1 is given by:

$$N_{k+1} - N_{k+1}P_{k+1} = (N_k - N_kP_k)(1 - I_k)(1 - q_k). \tag{9}$$

Independent exponential distributions are assumed for periods from the beginning of the age interval k to: (a) death in the absence of disease; (b) occurrence of disease; and (c) death in the presence of disease. The corresponding parameters are λ_1, λ_2, and λ_3. Given this model, λ_2 corresponds to the disease incidence rate ι_k (Podgor and Leske [1986]).

Transforming equations (4) and (9) results in the following:

$$\frac{1 - P_k}{1 - P_{k+1}} P_{k+1} e^{-(\lambda_1+\lambda_2)} = P_k e^{-\lambda_3} + (1 - P_k)(e^{-\lambda_3} - e^{-(\lambda_1+\lambda_2)}) \frac{\lambda_2}{\lambda_1 + \lambda_2 - \lambda_3}. \tag{10}$$

If λ_1 and λ_3 are known, λ_2 can be estimated for age interval k using the estimates for P_k and P_{k+1}. Equation (10) is non-linear in λ_2, so it can be solved by standard methods such as the Newton-Raphson algorithm. Estimates for the variance of $\hat{\lambda}_2$ are derived by the delta method.

If $\lambda_1 = \lambda_3$, and using $I_k = 1 - e^{-\lambda_2}$, equation (5) is obtained:

$$I_k = \frac{P_{k+1} - P_k}{1 - P_k},$$

as well as:

$$\lambda_2 = \iota_k = -\ln\left(\frac{P_{k+1} - P_k}{1 - P_k}\right). \qquad (11)$$

The cumulative incidence I_k corresponds to the crude probability for developing the disease since it disregards mortality risk. If mortality risk is considered, the net probability for developing the disease, R_k, becomes:

$$R_k = \frac{\lambda_2}{\lambda_1 + \lambda_2}(1 - e^{-(\lambda_1+\lambda_2)}). \qquad (12)$$

Applications of this method in the context of studies about eye diseases in England and HIV in Burundi can be found in publications of Podgor and Leske (1986) and Saidel et al. (1996), respectively.

2.2.3 Method of Elandt-Johnson and Johnson

For the case of equal mortality for affected and unaffected individuals, Elandt-Johnson and Johnson (1980) have developed a recursive formula for the calculation of the cumulative incidence in each age interval k using prevalence data:

$$I_{k+1} = 2(P_{k+1} - P_k) - I_k, \text{ where } I_0 = 2\,P_0. \qquad (13)$$

This implies that prevalence is taken at the midpoint of each interval. Unlike the methods of Leske et al. (1981), periods from the beginning of an age interval until occurrence of disease are assumed to be distributed uniformly when using the Elandt-Johnson and Johnson method. Also, neither an estimator for the variance of $\hat{I}_k$, nor a formula for ι_k is given. As above, the disease is supposed to be irreversible. Moreover, the disease must not be an uncommon one, otherwise estimates for the cumulative incidence could be negative.

Elandt-Johnson and Johnson have presented a similar method for the case of different mortality of affected and unaffected individuals, which requires data from life tables for affected individuals. However, these data are hardly available.

2.3 Methods of Gregson et al.

Gregson et al. (1996) present two different approaches for the estimation of the HIV-incidence with the use of seroprevalence data. In both cases the incidence rate is modeled as a continuous function of age. Age-specific incidence rates and HIV-related mortality are required to remain constant over time.

Hence, cumulative incidence as a function of age is given as follows:

$$\mathrm{I(a)} = \int_0^{\mathrm{a}} \iota(\mathrm{x})\mathrm{e}^{-\int_0^{\mathrm{x}} \iota(\mathrm{y})\mathrm{dy}}\mathrm{dx}. \tag{14}$$

2.3.1 Cumulative Incidence and Survival Method

The first of the two approaches of Gregson et al. (1996) is the cumulative incidence and survival method. Its objective is to interpret the prevalence at age a, P(a), as the cumulative incidence up to age a, adjusted for mortality. This means that the proportion of HIV-infected individuals with age a in the whole population is equivalent to the proportion of individuals infected at age a-x, x > 0, who survived the following x years. Thus, P(a) can be written as follows:

$$\mathrm{P(a)} = \frac{\int_0^{\mathrm{a}} \iota(\mathrm{x})\mathrm{e}^{-\int_0^{\mathrm{x}} \iota(\mathrm{y})\mathrm{dy}}\mathrm{p}^{+}_{\mathrm{x,a}}\mathrm{dx}}{\int_0^{\mathrm{a}} \iota(\mathrm{x})\mathrm{e}^{-\int_0^{\mathrm{x}} \iota(\mathrm{y})\mathrm{dy}}\mathrm{p}^{+}_{\mathrm{x,a}}\mathrm{dx} + \mathrm{e}^{-\int_0^{\mathrm{a}} \iota(\mathrm{y})\mathrm{dy}}}, \tag{15}$$

where both $\iota(\mathrm{a})$ and $\mathrm{p}^{+}_{\mathrm{x,a}}$ must be specified. Gregson et al. (1996) suggest the following age-specific pattern for the incidence rate:

$$\iota(\mathrm{a}) = \alpha \exp\left(-\frac{(\mathrm{a} - U\gamma)^2}{\beta(\frac{U}{2})^2}\right) \mathrm{I}_{[L,U]}(\mathrm{a}), \quad \alpha, \beta, \gamma > 0, \tag{16}$$

where L and U are the lower and upper age limits beyond which HIV-incidence is negligible. High values of the parameters α, β and γ tend to indicate high levels of incidence rate, similar level of incidence rate across a broad range of ages, and an older peak age-at-infection, respectively. The survival periods are assumed to follow a Weibull distribution with parameters φ and M, where M is the median survival period.

Using the estimates for P(a) and the maximum likelihood method, $\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\varphi}, \hat{\mathrm{M}}$ can be calculated and hence $\hat{\iota}(\mathrm{a})$. A variance estimation for the estimators of the parameters is not given. Confidence intervals are calculated by means of the log-likelihood ratio.

A similar approach can be found in an article by Ades and Nokes (1993). The incidence rate, $\iota(\mathrm{a,t})$, is modeled as a function of both age and time. Two different specifications of $\iota(\mathrm{a,t})$ are given. In order to separate the dependence on age from the dependence on time, serial prevalence studies are required.

2.3.2 Constant Prevalence Method

This method is similar to the one by Podgor and Leske (1986) in case of different mortality for affected and unaffected individuals. The cumulative incidence in the time interval [t-r, t] for the age group [a-r, a] is the net increase in prevalence within this cohort over the time interval, after adjusting for deaths to infected and uninfected individuals. However, whereas Podgor and Leske (1986) assume exponentially distributed survival periods (see 2.2.2), Gregson et al. (1996) estimate the mortality for each time interval.

After various transformations, the estimator for the incidence as function of age at time t results in:

$$\hat{\iota}_{\mathrm{t}}(\mathrm{a}) = \frac{\hat{\mathrm{P}}_{\mathrm{t}}(\mathrm{a,a+r})\left[1-\hat{\mathrm{P}}_{\mathrm{t}}(\mathrm{a\text{-}r,a})\right]+\hat{\mathrm{P}}_{\mathrm{t}}(\mathrm{a\text{-}r,a})\left[\hat{\mathrm{P}}_{\mathrm{t}}(\mathrm{a,a+r})-1\right]\pi^{+}_{\mathrm{a\text{-}r/2,a+r/2}}}{\left[1-\hat{\mathrm{P}}_{\mathrm{t}}(\mathrm{a\text{-}r,a})\right]\left[1-\hat{\mathrm{P}}_{\mathrm{t}}(\mathrm{a,a+r})\right]\pi^{+}_{\mathrm{a,a+r/2}}+\hat{\mathrm{P}}_{\mathrm{t}}(\mathrm{a,a+r})}. \tag{17}$$

It is assumed that infection with HIV occurs at the midpoints of the age and time intervals. The survival probability π^{+} is given as follows:

$$\pi^{+}_{\mathrm{a\text{-}r/2,a+r/2}} = \frac{\int_0^{\mathrm{a}} \iota(\mathrm{x})\mathrm{e}^{-\int_0^{\mathrm{x}}\iota(\mathrm{y})\mathrm{dy}}\,\mathrm{p}^{+}_{\mathrm{x,a+r/2}}\mathrm{dx}}{\int_0^{\mathrm{a}} \iota(\mathrm{x})\mathrm{e}^{-\int_0^{\mathrm{x}}\iota(\mathrm{y})\mathrm{dy}}\,\mathrm{p}^{+}_{\mathrm{x,a\text{-}r/2}}\mathrm{dx}}. \tag{18}$$

Consequently, the survival probabilities depend on the incidence rate, $\iota(\mathrm{a})$, which itself is a function of values of π^{+}. Thus, an iterative estimation procedure must be adopted choosing initial values for $\iota(\mathrm{a})$ for each age interval. Estimation of the variance of $\hat{\iota}(\mathrm{a})$ is done by means of the delta method.

Gregson et al. (1996) apply both methods to data comprising 3600 pregnant women in Uganda who were tested for HIV between 1989 and 1990 in several hospitals of the capital, Kampala. Incidence estimation is done for age intervals of three years' length. This implies that the discrete equivalents for the equations given above are used. Similar values for $\hat{\iota}(\mathrm{a})$ are obtained in both cases. The shape of the age-specific distribution is smoother in case of the cumulative incidence and survival method resulting from the more complete use of empirical data, on one hand, and the mathematical function assumed, on the other. When applying the constant prevalence method, empirical data of only one age interval are used for the estimation of $\iota(\mathrm{a})$.

2.4 Compartment Model of Keiding

Another approach to the estimation of incidence of a disease from prevalence data is presented by Keiding (1991). Keiding uses a compartment model composed of three states, 'healthy', 'ill', and 'dead', and the corresponding transition rates ι, μ, and ν. The disease is assumed to be chronic. The incidence rate, ι, corresponds to the intensity of the transition from 'healthy' to 'ill', whereas μ and ν denote the transition intensities from 'healthy' to 'dead' and 'ill' to 'dead', respectively.

Keiding describes these transitions with respect to dependence on age, time and duration of disease, considering different situations. One of these situations is the case of time independence in which information about the onset of the disease is not given. This situation corresponds to the reality of many seroprevalence studies in many countries in Africa and Asia where the actual date of infection by HIV is usually unknown.

The compartment model makes the following assumptions: births follow a Poisson process with parameter β; life histories of the individuals are independent of each other; and life histories are independent of the birth process. Consequently, the process of falling ill follows a Poisson process with age dependent intensity as follows (Brillinger [1986]):

$$\gamma(\mathrm{a}) = \beta \exp\left(-\int_0^{\mathrm{a}} (\mu(\mathrm{u}) + \iota(\mathrm{u}))\mathrm{du} \right) \iota(a). \tag{19}$$

Further, the joint probability density for an individual to fall ill at age y and survive until age a, a > y, is given by:

$$\mathrm{h}(\mathrm{y}, \mathrm{a}) = \exp\left(-\int_0^{\mathrm{y}} (\mu(\mathrm{u}) + \iota(\mathrm{u}))\,\mathrm{du} \right) \iota(y) \exp\left(-\int_{\mathrm{y}}^{\mathrm{a}} \nu(\mathrm{u})\mathrm{du} \right). \tag{20}$$

Thus, the expectation value, $\mathrm{E}(\mathrm{P_N}; \mathrm{a}, \mathrm{y})$, of the number of individuals with age $[\mathrm{a}, \mathrm{a}+\delta\mathrm{a}]$ after falling ill at age $[\mathrm{y}, \mathrm{y}+\delta\mathrm{y}]$ is:

$$\mathrm{E}(\mathrm{P_N};\, \mathrm{a}, \mathrm{y}) = \beta\, \mathrm{h}(\mathrm{y}, \mathrm{a})\, \delta\mathrm{y}\, \delta\mathrm{a}. \tag{21}$$

The net life time probability of falling ill becomes:

$$\mathrm{R} = \int_0^{\infty} \exp\left(-\int_0^{\mathrm{y}} (\mu(\mathrm{u}) + \iota(\mathrm{u}))\,\mathrm{du} \right) \iota(y)\mathrm{dy}. \tag{22}$$

Let $\Delta = \mathrm{E}(\mathrm{A}-\mathrm{Y}|\mathrm{A}>\mathrm{Y})$ be the conditional expectation value of duration of disease given that the individual has fallen ill. Then equation (2) of subsection 2.1 corresponds to:

$$\beta \int_0^\infty \int_y^\infty \mathrm{h}(\mathrm{y},\mathrm{a})\mathrm{da}\,\mathrm{dy} = \beta\,\mathrm{R}\,\Delta. \tag{23}$$

In order to estimate the age-specific incidence rate, nonparametric methods of survival analysis are used. For example, let G(a) represent the gross probability of contracting the disease before age a. That is:

$$\mathrm{G}(\mathrm{a}) = 1 - \exp\left(\int_0^{\mathrm{a}} \iota(\mathrm{u})\mathrm{du}\right) = \int_0^{\mathrm{a}} \iota(\mathrm{u})\mathrm{e}^{-\int_0^{\mathrm{u}} \iota(\mathrm{s})\mathrm{ds}}\mathrm{du}. \tag{24}$$

Thus, the likelihood function depending on ι and μ in case of equal mortality of affected and unaffected individuals has the following form:

$$\mathrm{L}(\iota,\,\mu) = \mathrm{L}(\mu)\mathrm{L}(\iota) = \mathrm{M}^{-\mathrm{n}} \prod_{i=1}^{\mathrm{n}} \exp\left(-\int_0^{\mathrm{a}_i} \mu(\mathrm{u})\mathrm{du}\right) \prod_{i=1}^{\mathrm{n}} (1-\mathrm{G}(\mathrm{a}_i))^{1-\delta_i}\mathrm{G}(\mathrm{a}_i)^{\delta_i}, \tag{25}$$

where M denotes the expected life time, δ_i the indicator function, whether an individual is sick or not, and a_i the age of individual i.

Maximizing the likelihood function gives a left-continuous estimator of G(a), G(a-), through which $\iota(\mathrm{a})$ can be estimated (Keiding [1991]). Using a kernel, K, with bandwidth b prevents the subdivision in several age intervals:

$$\hat{\iota}(\mathrm{a}) = \frac{1}{\mathrm{b}} \int \mathrm{K}\left(\frac{\mathrm{a}-\mathrm{u}}{\mathrm{b}}\right) \frac{\mathrm{d}\hat{\mathrm{G}}(\mathrm{u})}{1-\hat{\mathrm{G}}(\mathrm{u}-)}. \tag{26}$$

Keiding presents an application of this method for data on the incidence rate of Hepatitis A in Bulgaria.

In case of different mortality for affected and unaffected individuals, the estimation procedure for $\hat{\iota}(\mathrm{a})$ remains the same for the following assumptions: mortality of both affected and unaffected is known; the difference in mortality of affected and unaffected is small; and the disease of interest is rare. However, regarding the HIV-epidemic in many parts of Africa and Asia these assumptions are hardly fulfilled, especially in Subsaharan Africa and in Southeast Asia.

Brunet and Struchiner (1996) have similarly developed a compartment model in which they consider two groups with the same states, one group receiving a vaccine and the other group no vaccine. Transition intensities are modeled based on dependence on age and time and effects of a new treatment or vaccine could then be examined. However, serial prevalence studies are required to estimate transition rates depending on time.

3 Discussion

Cohort and serial seroprevalence studies such as those conducted in Ruanda (Bucyendore et al. [1993]) and Burundi (Saidel et al. [1996]), are still the exception in many parts of Africa. Usually, single seroprevalence studies are conducted for the estimation of HIV-incidence. This implies that conclusions about the dynamics of the spread of the HIV-epidemic can only be drawn from data which reflect the situation at one single moment.

Brookmeyer and Gail (1994; Chap. 3) and Brundage et al. (1990) discuss general problems of seroprevalence studies with emphasis on potential sources of bias. Possible sources include tests for HIV with low sensitivity or specifity, which might be a big problem where there is low HIV-incidence.

3.1 Practical Considerations of the Methods

Most of the methods discussed in Section 2 could be used to estimate HIV-incidence. However, the methods by Leske et al. (1981) and Keiding (1991) for the case of equal mortality of affected and unaffected individuals do not seem to be appropriate due to the high mortality of AIDS.

Due to the fact that the approach of Freeman and Hutchison (1980) does not model incidence as a function of age, it is not appropriate for the case of infection with HIV. However, the main equation, 'Prevalence = Incidence $\times$ Duration of Disease', can be applied for different age intervals. For this, the average life expectancy with HIV would be necessary to be available for each age interval. A constant number of affected and unaffected individuals over time is required since it is normally not given at the beginning of an epidemic. A generalization of this method for the case of exponentially growing populations is presented by Alho (1992).

Brookmeyer and Quinn (1995) refer to several sources of estimation bias in their study about the HIV-incidence in India using the equation of Freeman and Hutchison (1980). The main source of uncertainty is the estimation of the duration of the p24 antigen period. Due to these uncertainties the method seems to be unsuitable in connection with the HIV-epidemic. Large sample sizes would be required to estimate the expectation value of the duration of the p24 antigen period with low variance, especially if dependence on age was modeled. In the context of another study, Brookmeyer et al. (1995) combine the method of Freeman and Hutchison (1980) with the simple estimator for cumulative incidence in case of cohort studies. The purpose is to eliminate the main sources of uncertainty in both methods.

In connection with several seroprevalence studies in Burundi, Saidel et al. (1996) compare the incidence estimations with the use of the method of Podgor and Leske (1986) and with methods which utilize the prevalence values of these studies. They report that incidence estimation coincides for the chosen age intervals. This indicates that the method of Podgor and Leske is quite exact. This result is restricted to situations in which the HIV-epidemic has reached a stable level. As in other methods, the more stable an epidemic is, the more exact are the incidence estimations.

In contrast with Saidel et al. (1996), Ades and Nokes (1993) show that incidence estimations are biased when dependence on time is not modeled. They present results of a study about the spreading of Toxoplasmose in South Yorkshire, England. In order to model separately dependence on age and on time, data are required for at least two moments, unless additional information is given, for example, with regard to date of infection. In many parts of Asia and Africa where HIV is mainly transmitted via heterosexual relations, this information is normally not available. Thus, a steady-state population verifiable by studies over time must be assumed.

The constant prevalence method of Gregson et al. (1996) is similar to the method of Leske et al. (1981). However, unlike the latter, negative values for the estimation of the incidence rate are possible. It should be noted that both methods require constant age-specific incidence over time. Comparing the constant prevalence method with the cumulative incidence and survival method, the condition of constant age-specific incidence over time is more restrictive for the latter. This is because the estimation of the incidence density for the age interval k requires prevalence estimates of all previous age intervals. In case of the constant prevalence method, only the prevalence estimate of the previous age interval is needed. This leads to less variation for the incidence estimates when the cumulative incidence and survival method is used. Less variation using the cumulative incidence and survival method is also due to the specification of the incidence density depending on the age of the individual. In this case therefore, estimation accuracy depends on the specification of the incidence rate.

The accuracy of the nonparametric method of survival analysis, as presented by Keiding (1991), depends on whether the assumption that births follow a Poisson process is justified. Besides, the chosen kernel estimate determines the form of the age specific incidence rate.

3.2 Future Possibilities

A comparison of the methods as presented above is needed to determine which method is the most appropriate one in a given situation. This comparison could be done theoretically by examining the expectation values and variances of the different estimators and in a practical way using simulations or real data. See for example Saidel et al. (1996).

In addition, it is necessary to obtain more information about the population while conducting seroprevalence studies. Gregson et al. (1996) suggest that a statistical survey of the potential migration within the population should be made since the assumed steady-state population could be affected by migration. Regarding epidemics caused by parasites, de Vlas et al. (1993) model disease incidence in dependence on the amount of parasites in the blood of the individual. If the possibility exists to obtain the virusload while testing for HIV, the method of de Vlas et al. might be adaptable to the HIV-epidemic.

By means of the virusload, conclusions about the time of infection could be drawn. Consequently, the second step could be a more exact estimation of the present incidence rate. A similar model for the estimation of the number of AIDS cases, as used by the World Health Organization, is presented by Chin and Lwanga (1991). The backward calculation method used by the WHO can be used even when the virusload is unknown since it is the incubation period distribution that is to be modeled.

Furthermore, Bayesian methods could be developed in order to utilize any prior information about the incidence rate. This might allow for more accurate estimations.

Finally, further investigations could be focused on the geographical spread of the HIV-epidemic using methods of spatial statistics. Especially in countries having one big urban centre, the capital, with surrounding rural provinces, there might be a big difference between the spread of HIV-infection in urban and rural communities. On the other hand, due to migration from and to the urban centre, the dynamic of the spread of HIV in urban and rural areas will be interrelated. Examining the spatial spread of the epidemic over time might aid the planning of efficient HIV-prevention campaigns in the future.

References

Ades, A. and D. Nokes (1993). Modeling age- and time-specific incidence from seroprevalence: Toxoplasmosis. *American Journal of Epidemiology* **137**(9), 1022–34.

Alho, J. (1992). On prevalence, incidence, and duration in general stable populations. *Biometrics* **48**, 587–92.

Brillinger, D. (1986). The natural variability of vital rates and associated statistics. *Biometrics* **42**, 693–734.

Brookmeyer, R. and M. Gail (1994). *AIDS epidemiology: a quantitative approach.* New York: Oxford University Press.

Brookmeyer, R. and T. Quinn (1995). Estimation of current Human Immunodeficiency Virus incidence rates from cross-sectional survey using early diagnostic tests. *American Journal of Epidemiology* **141**(2), 166–72.

Brookmeyer, R., T. Quinn, M. Shepherd, S. Mehendale, J. Rodrigues, and R. Bollinger (1995). The Aids epidemic in India: A new method for estimating current Human Immunodeficiency Virus (HIV) incidence rates. *American Journal of Epidemiology* **142**(7), 709–13.

Brundage, J., D. Burke, L. Gardner, J. McNeil, M. Goldenbaum, R. Visintine, R. Redfield, M. Peterson, and R. Miller (1990). Tracking the spread of the HIV infection epidemic among young adults in the United States: Results of the first four years of screening among civilian applicants for U.S. military service. *Journal of Acquired Immune Deficiency Syndromes* **3**, 1168–80.

Brunet, R. and C. Struchiner (1996). Rate estimation from prevalence information on a simple epidemiologic model for health interventions. *Theoretical Population Biology* **50**, 209–26.

Bucyendore, A., P. van de Perre, E. Karita, A. Nziyumvira, I. Sow, and E. Fox (1993). Estimating the seroincidence of HIV-1 in the general adult population in Kigali, Rwanda. *AIDS* **7**, 275–77.

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

de Vlas, S., N. Nagelkerke, J. Habbema, and G. van Oortmarssen (1993). Statistical models for estimating prevalence and incidence of parasitic diseases. *Statistical Methods in Medical Research* **2**, 3–21.

Ederer, F. (1964). The effect of adjusting for competing mortality risks. *American Journal of Public Health* **54**(7), 1129–33.

Elandt-Johnson, R. and N. Johnson (1980). *Survival models and data analysis.* New York: Wiley.

Freeman, J. and G. Hutchison (1980). Prevalence, incidence and duration. *American Journal of Epidemiology* **112**(5), 707–23.

Gregson, S., C. Donnelly, C. Parker, and R. Anderson (1996). Demographic approaches to the estimation of incidence of HIV-1 infection among adults from age-specific prevalence data in stable endemic conditions. *AIDS* **10**, 1689–97.

Heyward, W., S. Osmanov, J. Saba, J. Esparza, E. Belsey, R. Stoneburner, J. Kaldor, and P. Smith (1994). Preparation for Phase III HIV vaccine efficacy trials: methods for the determination of HIV incidence. *AIDS* **8**, 1285–91.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society* **154**(3), 371–412.

Kreienbrock, L. and S. Schach (1995). *Epidemiologische Methoden.* Stuttgart · Jena: Gustav Fischer Verlag.

Leske, M., F. Ederer, and M. Podgor (1981). Estimating incidence from age-specific prevalence in Glaucoma. *American Journal of Epidemiology* **113**(5), 606–13.

Podgor, M. and M. Leske (1986). Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. *Statistics in Medicine* **5**, 573–78.

Podgor, M., M. Leske, and F. Ederer (1983). Incidence estimates for lens changes, macular changes, open-angle Glaucoma and diabetic retinopathy. *American Journal of Epidemiology* **118**(2), 206–12.

Saidel, T., D. Sokal, J. Rice, T. Buzingo, and S. Hassig (1996). Validation of a method to estimate age-specific Human Immunodeficiency Virus (HIV) incidence rates in developing countries using population-based seroprevalence data. *American Journal of Epidemiology* **144**(3), 214–23.