*J. R. Statist. Soc.* B (1999)
**61**, *Part* 2, *pp.* 287–307

# Statistical studies of infectious disease incidence

Niels G. Becker

*La Trobe University, Bundoora, Australia*

and Tom Britton

*Uppsala University, Sweden*

[Received September 1997. Final revision June 1998]

**Summary.** Methods for the analysis of data on the incidence of an infectious disease are reviewed, with an emphasis on important objectives that such analyses should address and identifying areas where further work is required. Recent statistical work has adapted methods for constructing estimating functions from martingale theory, methods of data augmentation and methods developed for studying the human immunodeficiency virus–acquired immune deficiency syndrome epidemic. Infectious disease data seem particularly suited to analysis by Markov chain Monte Carlo methods. Epidemic modellers have recently made substantial progress in allowing for community structure and heterogeneity among individuals when studying the requirements for preventing major epidemics. This has stimulated interest in making statistical inferences about crucial parameters from infectious disease data for such community settings.

*Keywords*: Control of diseases; Data augmentation; Epidemic data; Epidemics in structured communities; Heterogeneous individuals; Immunity coverage; Incomplete data; Infectivity function; Martingale estimating function; Vaccination coverage; Vaccine efficacy

## 1. Introduction

This paper was prepared for the Royal Statistical Society Epidemics Workshop, held on the Isle of Skye, Scotland, on March 31st–April 12th, 1997. Our purpose is to identify statistical issues and problems that are current and worthwhile as projects for research collaboration. The hope that collaboration would begin on one or two of these problems at the workshop has been fulfilled, but there remain numerous pressing problems. An attempt is made to cast a wide net and to be objective, but it is inevitable that the problems identified draw heavily on the authors' own experience. The paper includes a requisite review of material on epidemic models and the estimation of their parameters.

Infectious disease data have two features that distinguish them from other data. They are highly dependent and the infection process is only partially observable. A consequence of these features is that the analysis of data is usually most effective when it is based on a model that describes aspects of the infection process, i.e. on a transmission model. Therefore modelling is an integral part of statistical work in this area. Although deterministic models can be a guide towards parameter estimates, the need to quantify the precision of estimates and the variation in data imply that stochastic models are the natural basis for the analysis of infectious disease data.

*Address for correspondence*: Niels G. Becker, Department of Statistical Science, La Trobe University, Bundoora, Victoria 3083, Australia.
E-mail: n.becker@latrobe.edu.au

 1369–7412/99/61287

Statistical analysis, embracing modelling, parameter estimation, hypothesis testing and the design of studies, plays an essential role in bridging the gap between the mathematical theory and public health practice, and it is this aspect that motivates the present discussion. In other words, we attempt to promote the use of statistical analyses that provide practical insight and guidance for disease control, with emphasis on identifying issues that have not been addressed adequately.

The paper is primarily concerned with the incidence of a disease that is transmitted by person-to-person contact, Section 3.1 being an exception. Modelling and analysis of the development of the disease within a host is not covered; see Isham and Medley (1989), part 2, for a collection of recent papers on this topic. Analyses concerned with seroepidemiological data, such as seroprevalence, are not covered either; see Brookmeyer *et al.* (1995) and Saidel *et al.* (1996) for recent examples of such work.

Section 2 provides an overview of some standard epidemic models and methods for making statistical inference about their parameters. Its aims are both to help to make the paper self-contained and to act as a foundation for the discussion in the rest of the paper. Later sections are directed at specific objectives. Methodology that can help to determine the way that the disease is transmitted is reviewed in Section 3. In Section 4 the focus is on identifying sources of heterogeneity in disease transmission, as well as on estimation problems when heterogeneity is present. The particular heterogeneity arising from having a structured community, and assessing the consequences of ignoring such heterogeneity, is investigated in greater detail in Section 5. A major motivation for the study of epidemic models is the insight that they provide about the control of disease transmission. Section 6 looks at the estimation of parameters that are needed to determine control specifications, such as the vaccination coverage required to prevent epidemics and the estimation of vaccine efficacy. A relatively new area of work is that concerned with the transmission of the human immunodeficiency virus (HIV), the virus that leads to acquired immune deficiency syndrome (AIDS), which has some unique features. In Section 7 we focus on some of this methodology to see whether these methods can be utilized for the analysis of data on other infectious diseases. In the final section we identify some statistical problems that are worthy of further attention.

## 2. Parameter estimation for standard epidemic models

### 2.1. Susceptible–infective–removed models in continuous time

For many transmissible diseases all individuals are initially susceptible. On infection they become infectious for a period, after which they stop being infectious, recover and become immune. They are then said to be removed. An individual who is infectious is called infective. For convenience, models which assume that individuals pass, in turn, through the *susceptible*, *infective* and *removed* states are called SIR models.

#### 2.1.1. The general epidemic

The so-called general epidemic model was first studied by McKendrick (1926). It is a Markovian, continuous time model describing the spread of an SIR infectious disease in a population of homogeneous individuals who mix uniformly.

Consider a closed community with $n$ individuals. Let $S(t)$, $I(t)$ and $R(t)$ respectively denote the number of susceptible, infectious and removed individuals at time $t$. The relationship $R(t) = n - S(t) - I(t)$ holds for all $t$. Let $N(t) = S(0) - S(t)$ be the number of individuals infected in $(0, t]$. The initial conditions of an epidemic are specified by $S(0) = s_0$, $I(0) = i_0$ and

$R(0) = r_0$. In this section we assume these quantities to be known. In practice, the number of individuals who are immune is often unknown and this is in itself a source of challenging statistical problems; see Section 6.1.

The two processes $N(t)$ and $R(t)$ are increasing counting processes. Let $\mathcal{H}_t$ denote the $\sigma$-algebra generated by the history $\{S(u), I(u); 0 \leqslant u \leqslant t\}$. Then the general epidemic is defined by

$$\left.\begin{aligned} \Pr\{\mathrm{d}N(t) = 1, \mathrm{d}R(t) = 0 \,|\, \mathcal{H}_t\} &= \beta \bar{S}(t) I(t)\,\mathrm{d}t + o(\mathrm{d}t), \\ \Pr\{\mathrm{d}N(t) = 0, \mathrm{d}R(t) = 1 \,|\, \mathcal{H}_t\} &= \gamma I(t)\,\mathrm{d}t + o(\mathrm{d}t), \\ \Pr\{\mathrm{d}N(t) = 0, \mathrm{d}R(t) = 0 \,|\, \mathcal{H}_t\} &= 1 - \beta \bar{S}(t) I(t)\,\mathrm{d}t - \gamma I(t)\,\mathrm{d}t + o(\mathrm{d}t), \end{aligned}\right\} \tag{1}$$

where $\bar{S}(t) = S(t)/n$ is the proportion susceptible at $t$ (this *overline* notation will also be used for other proportions). After some finite (random) time $\tau$ no infectious individuals are present in the community and the epidemic is over. The number of individuals who are still susceptible at time $\tau$ specifies the *final state* of the epidemic.

The general epidemic has two parameters, namely $\beta$ and $\gamma$. The parameter $\beta$ is the rate with which an infectious individual has close contacts with other members of the community, $\beta \bar{S}(t)$ is the rate of close contacts with susceptible individuals and so $\beta \bar{S}(t) I(t)$ is the aggregated rate at which infectious individuals have close contacts with susceptible individuals. In statistical studies it is natural to transform the other parameter to $\gamma^{-1}$, the mean duration of the infectious period; the model implicitly assumes that the infectious period is exponentially distributed with parameter $\gamma$. The quantity $\theta = \beta/\gamma$ is the average number of infections caused by one infectious individual during the early stages of the epidemic. It is called the basic reproduction number, often denoted $R_0$, and plays an important role in SIR models, e.g. Ball (1983). For a large class of epidemic models it is known that two qualitatively different situations may occur with an epidemic in a large population. Either the epidemic dies out quickly, with a small number of individuals infected, or the epidemic takes off and a positive fraction (with some Gaussian noise) of the community become infected before the epidemic dies out. The latter case, called a *major outbreak*, can only happen if $R_0 = \theta > 1$, e.g. Ball (1983).

The general epidemic model has several features that are open to criticism. The assumption of homogeneous individuals is relaxed in Section 4, whereas that of uniform mixing is relaxed in Section 5. The length of the infectious period may follow a different distribution. Indeed, the force of infection exerted might vary during the infectious period, in which case the length of the infectious period assumes a less important status in the analysis. In particular, some diseases are known to have a latent period following infection, during which the individual is not infectious. Some resulting statistical issues are discussed briefly in Section 3.2.

### 2.1.2. *Maximum likelihood estimation under complete observation*

Assume that a realization of the general epidemic is observed completely and continuously up to the end of the epidemic. Using counting process theory (e.g. Andersen *et al.* (1993), p. 402), we may then write the log-likelihood explicitly as

$$l(\beta, \gamma) = \int_0^\tau [\log\{\beta \bar{S}(u) I(u)\}\,\mathrm{d}N(u) - \beta \bar{S}(u) I(u)\,\mathrm{d}u + \log\{\gamma I(u)\}\,\mathrm{d}R(u) - \gamma I(u)\,\mathrm{d}u].$$

It is easily verified that the maximum likelihood (ML) estimators are given by

$$\hat{\beta} = N(\tau) \Big/ \int_0^\tau \bar{S}(u)\, I(u)\, \mathrm{d}u,$$
$$\widehat{\gamma^{-1}} = \int_0^\tau I(u)\, \mathrm{d}u \Big/ \{R(\tau) - r_0\}. \tag{2}$$

Note that $\int_0^\tau I(u)\, \mathrm{d}u$ is the sum of all infectious periods, so $\widehat{\gamma^{-1}}$ is simply the sample mean of the infectious periods.

Rida (1991) showed that, on the part of the sample space where a major epidemic occurs, the ML estimators are consistent and asymptotically independent normal variates as $n$, the population size, tends to $\infty$. The standard errors of the estimators are then given by

$$\mathrm{se}(\hat{\beta}) = \hat{\beta}/\sqrt{N(\tau)}$$

and

$$\mathrm{se}(\widehat{\gamma^{-1}}) = \widehat{\gamma^{-1}}/\sqrt{\{R(\tau) - r_0\}}.$$

Under complete observation several generalizations of the present model can be incorporated without meeting major difficulties in estimation. For example it is possible to allow for a latency period, an arbitrary distribution of the infectious period, that the infection rate depends on time, $\beta = \beta(t)$, perhaps due to seasonal changes, and different types of individual.

Often the epidemic is observed only partially. This makes ML estimation cumbersome, and other approaches to statistical inference become attractive.

### *2.1.3. Martingale methods under incomplete observation*

It might be that only the final state of the epidemic is observed, i.e. besides the initial state ($s_0$, $i_0$, $r_0$) the data consist of $S(\tau)$, which determines $R(\tau) = n - S(\tau)$ since $I(\tau) = 0$. Estimation procedures for this type of data have been proposed by several researchers, e.g. Becker (1989), section 7.4.1. Another plausible data set consists of the initial values and continuous observation of the removal process. The final state of the epidemic process can then be deduced, so this is a more comprehensive data set. The time of removal can for many diseases be approximated by the time of diagnosis or show of symptoms, which is sometimes available, whereas the time of infection is rarely known. Estimation for this type of data is treated by Bailey (1975), section 6.82, and more recently by Becker and Hasofer (1997, 1998).

When the epidemic is observed only partially the likelihood cannot be written in a closed form. Instead estimation can, for example, be based on approximations of certain recursive formulae defining the likelihood as in Bailey (1975), p. 118, or rely on large population approximations. Alternatively martingale techniques can be used with the method of moments to give estimates with explicit expressions. Two zero-mean $\mathcal{H}_t$-martingales are defined by

$$M_1(t) = N(t) - \int_0^t \beta\, \bar{S}(u)\, I(u)\, \mathrm{d}u, \tag{3}$$

$$M_2(t) = R(t) - r_0 - \int_0^t \gamma\, I(u)\, \mathrm{d}u. \tag{4}$$

From martingale theory it follows that

$$M(t) = \int_0^t \frac{1}{\bar{S}(u-)}\,\mathrm{d}M_1(u) - \frac{\beta}{\gamma}M_2(t) = \frac{n}{s_0} + \frac{n}{s_0 - 1} + \ldots + \frac{n}{S(t)+1} - \frac{\beta}{\gamma}\{R(t) - r_0\}$$

is a zero-mean martingale. By equating $M(\tau)$ to its mean we obtain the estimate

$$\hat{\theta} = \left\{ \frac{n}{s_0} + \frac{n}{s_0 - 1} + \ldots + \frac{n}{S(\tau)+1} \right\} \Big/ \{R(\tau) - r_0\} \approx -\log(1 - \tilde{p})/\{\bar{R}(\tau) - \bar{r}_0\}, \qquad (5)$$

where $\tilde{p} = 1 - S(\tau)/s_0$ is the observed proportion who became infected. Note that $\hat{\theta}$ depends only on the initial and final states of the epidemic. Applying the martingale central limit theorem it can be shown (see Rida (1991)) that for a major epidemic in a large community the estimator $\hat{\theta}$ is approximately normally distributed with mean $\theta$ and a standard deviation that is estimated consistently by

$$\mathrm{se}(\hat{\theta}) = \left[ \frac{1}{s_0^2} + \frac{1}{(s_0 - 1)^2} + \ldots + \frac{1}{\{S(\tau)+1\}^2} + \frac{\hat{\theta}^2}{n}\{\bar{R}(\tau) - \bar{r}_0\} \right]^{1/2} \Big/ \{\bar{R}(\tau) - \bar{r}_0\}.$$

We can only estimate $\theta$, or a function thereof, when only the final state is observed. This is not surprising since we observe one random quantity $S(\tau)$ and the distribution of this quantity is independent of the unit of time, as is $\theta = \beta/\gamma$. However, if the removal process is observed continuously we can do more. Becker and Hasofer (1997) applied martingale techniques to obtain another estimating equation, which enables $\beta$ and $\gamma$ to be estimated separately. The process that they considered is $\tilde{M}(t) = S(t)(1 + \theta/n)^{R(t)}$ which can be shown to be a martingale by using equations (3) and (4). From this martingale it is possible to construct an estimating equation depending only on the final state and the removal process and to derive estimates and confidence regions for the two parameters $\beta$ and $\gamma$.

We saw earlier that estimation is straightforward when the epidemic is fully observed. Thus, if it were possible to reconstruct the complete epidemic process from the available data we could estimate parameters in a simple way. One example of this idea is presented at the end of Section 7.2.

### 2.2. Epidemic chain models

In this section we consider analyses based on SIR models in discrete time. For some diseases the latent period is long in relation to the infectious period and neither period varies much between individuals. Chicken-pox, measles and mumps are considered to satisfy these criteria. In such cases it is sometimes possible to identify the *generation* in which an individual was infected. By generation is meant the number of predecessors in the chain tracing back to the introductory case(s). In applications it is only possible to distinguish the first few generations. For this reason direct applications of chain models occur mainly in statistical analyses of outbreaks in small groups, such as households. However, under some circumstances, e.g. under the Reed–Frost assumption described below, chain models are also a natural tool for deriving the distribution for the size of an outbreak irrespective of whether the separate generations can be distinguished.

#### 2.2.1. Chain–binomial models for independent household outbreaks

In a chain–binomial model the parameter $q_i$ is defined as the probability that a susceptible individual escapes infection when exposed to $i$ infectious individuals of a given generation. The events that different susceptible individuals escape infection are assumed to be independent.

Let $I_k$ denote the number of infective individuals in the $k$th generation and $S_k$ the number still susceptible. Then the probability function of the next generation is given by

$$\Pr(I_{k+1} = x | S_k = s, I_k = i; \mathbf{q}) = \binom{s}{x} p_i^x q_i^{s-x}, \qquad x = 0, \ldots, s, \tag{6}$$

where $\mathbf{q} = (q_1, q_2, \ldots, )$ and $p_i = 1 - q_i$. Individuals who are not infected remain susceptible the next generation, i.e. $S_{k+1} = S_k - I_{k+1}$, and individuals who are infectious become removed the next generation.

An epidemic chain specifies the spread through the generations. For example $1 \to 2 \to 1 \to 0$ means that $i_0 = 1$, $i_1 = 2$, $i_2 = 1$ and $i_3 = 0$. The initial numbers of infective and susceptible individuals, $(i_0, s_0)$, are assumed given. By sequential use of equation (6), the probability of a specific chain is

$$p_{s_0,i_0}(\mathbf{i}; \mathbf{q}) = \Pr(i_0 \to i_1 \to \ldots \to i_k | s_0, i_0; \mathbf{q}) = \frac{s_0!}{i_1!\, i_2! \ldots i_k!\, s_k!} \prod_{j=0}^{k-1} p_{i_j}^{i_{j+1}} q_{i_j}^{s_{j+1}}. \tag{7}$$

In equation (7) $i_k = 0$ and consequently $s_k = s_{k-1}$. Two specific parametric forms for $\mathbf{q}$ have received much attention in the literature: the Reed–Frost form, $q_i = q^i$, and the Greenwood form, $q_i = q$ $(i > 0)$. In the Reed–Frost case an individual must escape infection from each infected individual to remain susceptible, and this occurs independently. The Greenwood assumption says that the probability of escaping infection is constant as long as someone is infectious, an explanation being that the chance of infection depends only on whether or not the household is 'contaminated'.

Model (7) can be generalized in several ways. For example, the probability of escaping infection may depend on the type of individual, the specific generation and/or the household size. Another generalization is to assume that $\mathbf{q}$ is random with some specified distribution, independent and identically distributed for different individuals, resulting in a random effects model. More details of such models are given in Becker (1989).

### 2.2.2. *Epidemic chain data*

If the epidemic chain is observed at each generation the statistical analysis is straightforward, even with more general epidemic chain models. Assume that several different household outbreaks are observed and that the outbreaks may be treated as *independent*. Let $n(s_0, i_0, \ldots, i_k)$ denote the observed number of households with initial configuration $(s_0, i_0)$ and with epidemic chain $i_0 \to i_1 \to \ldots \to i_k$ and let $\mathbf{n} = \{n(s_0, i_0, \ldots, i_k)\}$. Then the log-likelihood is

$$l(\mathbf{q}; \mathbf{n}) = c + \sum_{s_0, i_0, \ldots, i_k} n(s_0, i_0, \ldots, i_k) \log\{p_{s_0 i_0}(\mathbf{i}; \mathbf{q})\}, \tag{8}$$

where $c$ is independent of the parameter vector $\mathbf{q}$.

The ML estimators $\{\hat{q}_i\}$ are obtained by maximizing the likelihood with respect to $\mathbf{q}$. Let $e_i$ denote the number of times that an individual escaped infection when exposed to $i$ infected the previous generation, and let $c_i$ denote the corresponding number who did not escape infection (note that these data are contained in the observed $\mathbf{n}$). Then the ML estimators are given by $\hat{q}_i = e_i/(e_i + c_i)$ when there are no restrictions on the parameters. The estimators are consistent and asymptotically independent normal variates (as the number of households becomes large) with a standard deviation that is consistently estimated by $\mathrm{se}(\hat{q}_i) = \{\hat{q}_i \hat{p}_i/(e_i + c_i)\}^{1/2}$.

ML inference for the single parameter $q$ under the Greenwood assumption ($q_i = q$) is similar to this, and straightforward. Under the Reed–Frost assumption ($q_i = q^i$) standard ways of making ML inferences are cumbersome when data on larger households are present, because the likelihood contains terms of the form $p_i^r = (1 - q^i)^r$. The EM algorithm facilitates such ML inferences. For application of the EM algorithm take the more detailed data set to be that which for each generation not only records whether a susceptible individual is infected but also records for each infected susceptible individual the number of infective individuals that contacted it. The event of being infected, which has probability

$$1 - q^i = \sum_{j=1}^{i} \binom{i}{j} p^j q^{i-j},$$

is then replaced by $i$ distinct events occurring with probabilities $\binom{i}{j} p^j q^{i-j}$, $j = 1, \ldots, i$. The maximization step of the EM algorithm is then immediate, because the likelihood for the detailed data has a form as for binomial data. The expectation step is also simple because its conditional expectation is determined by the mean of a binomial variate; see Dempster *et al.* (1977) and Becker (1997) for details.

### 2.2.3. *Size of outbreak data*

Often only the final number of infected individuals is observed, rather than the whole epidemic chain, i.e. for each household the data consist of $r = \Sigma_{k \geqslant 1} i_k$, besides $i_0$ and $s_0$. Let $m(r; i_0, s_0)$ denote the observed number of households with initial configuration $(i_0, s_0)$ which, besides the introductory case(s), eventually had $r$ infected individuals, and $\mathbf{m} = \{m(r; i_0, s_0)\}$. For these data the log-likelihood is given by

$$l(\mathbf{q}; \mathbf{m}) = c + \sum_{s_0, i_0, r} m(r; i_0, s_0) \log\{P_{s_0 i_0}(r; \mathbf{q})\},$$

where $P_{s_0 i_0}(r; \mathbf{q}) = \Sigma_{\mathbf{i}:|\mathbf{i}|=r} p_{s_0 i_0}(\mathbf{i}; \mathbf{q})$ and the latter are defined in equation (7). Direct maximization of this likelihood is cumbersome, since $P_{s_0 i_0}(r; \mathbf{q})$ is a sum of terms and so its logarithm has complicated derivatives.

Again the EM algorithm is useful. For the more detailed data set we *pretend* that the epidemic chains are observed. Begin with an initial estimate $\mathbf{q}^{(0)}$ of the parameter. The E-step is to calculate the conditional expectation of the log-likelihood for the complete data $\mathbf{N}$ given the observed data $\mathbf{m}$, giving

$$E\{l(\mathbf{q}; \mathbf{N}) | \mathbf{M} = \mathbf{m}; \mathbf{q}^{(0)}\} = \sum_{s_0, i_0, r} \frac{m(r; s_0, i_0)}{P_{s_0 i_0}(r; \mathbf{q}^{(0)})} \sum_{|\mathbf{i}|=r} p_{s_0 i_0}(\mathbf{i}; \mathbf{q}^{(0)}) \log\{p_{s_0 i_0}(\mathbf{i}; \mathbf{q})\}. \tag{9}$$

The M-step requires us to maximize this expression with respect to $\mathbf{q}$ to obtain a new estimate $\mathbf{q}^{(1)}$. This is simple since when we substitute equation (7) into equation (9) the latter takes on a form like that for binomial data. The E- and the M-steps are sequentially repeated, updating the estimate each time, until the increase in the log-likelihood at the new estimate value is negligible.

When observing the whole epidemic chain both $i_0$ and $s_0$ are automatically known, and not just their sum $i_0 + s_0 = h$, the household size. When the final size is observed for affected households the number of introductory cases may be unknown. One solution to this problem is to assume that in each household, independently of others, $i_0$ is a realization of a random variable conditioned to being strictly positive. Thus a conditional bin($h$, $\pi$) distribution truncated by excluding the zero class is appropriate. If uninfected households are also

observed $i_0$ is modelled simply as coming from a bin($h$, $\pi$) distribution. The same idea has been proposed for the analysis of other epidemic models; see Longini and Koopman (1982) and Addy *et al*. (1991). When applying the EM algorithm we now include observations on the $i_0$s in the complete data set and maximize with respect to $\mathbf{q}$ and the additional parameter $\pi$.

The chain–binomial model relies on two fundamental assumptions: that of independence between household outbreaks and that the disease evolves in generations. The assumption of independence is of course questionable, but very convenient. Fortunately, for a model explicitly allowing infections between households Ball *et al*. (1997) have shown that, given a major outbreak, household outbreaks are approximately (i.e. asymptotically) independent, if the number of households is large. If the randomness in the latent and infectious period is not negligible and/or the infectious period is not short relative to the latent period, then the chain–binomial model is not appropriate, at least in the case with separate $q_i$-parameters. Instead, we then need to resort to recursive formulae defining the distribution of the final size, e.g. Addy *et al*. (1991). Besides allowing latent and infectious periods to be random, it may be necessary to allow individuals to be of different type, e.g. according to age group and/or gender; see Section 4.

## 3. How is the disease transmitted?

### 3.1. Is it person-to-person transmission?

So far we have assumed that the disease spreads *locally* through person-to-person contacts. This is not always true. There may be a *global* source of infection, e.g. through a commonly shared water source or transmission via airborne organisms. Statistical analysis of data from such diseases differs in that there are no strong dependences arising from interaction between infective and susceptible individuals, making the likelihood tractable to work with. When analysing a disease with unknown means of spread it is therefore important to distinguish between the two types of spread.

When the spreading mechanism is local, cases will typically cluster according to some local community structure. The local structure which has received most attention in the literature is the presence of households, the reason being that members of the same household have a higher contact rate, but also because data containing household details are often available. Several different testing procedures are available to detect clustering of infected cases within households, usually based on final size data; see for example Britton (1997a) and references therein. Many test statistics are of the form

$$H = \sum_r f(h_r)\binom{N_r}{2} - g(h_r)N_r,$$

where $h_r$ denotes the size of household $r$, $f(h_r)$ and $g(h_r)$ are functions of the household size and $N_r$ is the (random) number of cases in household $r$ (so $\binom{N_r}{2}$ is the number of infected *pairs* of individuals in household $r$). Local spread tends to cluster cases in certain households, thus making the statistic $H$ large and the test significant. The distribution of $H$ is calculated under the null hypothesis of global spread and conditional on the overall proportion of cases in the population.

If the population consists of plants or animals, domestic or wild, some other local structure (e.g. geographical) might be more relevant than households. Let the local structure be specified by $\{d_{i,j}\}$ where $d_{i,j}$ is large if individual $j$ is 'close' to $i$ ($i \neq j$). For an SIR model with

arbitrary distribution of the infectious period the score statistic of the hypothesis that the local structure is irrelevant for the disease is then given by

$$T = \sum_{i,j} d_{i,j} C_i (C_j - \tilde{p}),$$

where $C_i$ indicates whether or not individual $i$ was a case and $\tilde{p}$ is the observed overall proportion infected (Britton, 1997b). Again, the distribution of $T$ is computed under the null hypothesis of uniform global spread and conditioned on $\tilde{p}$.

Suppose now that the spread is known to be local or some test indicates that it is local and we wish to find out more about the mechanism of spread. With chain–binomial models this can be done by exploring the dependence of the $q_i$, the probability of escaping infection when exposed to $i$ infected individuals of a given generation, on $i$ and other factors. For example, treating the spread within households, we might ask: does the risk of being infected increase with the number of infectious individuals? 'Yes' would indicate person-to-person transmission whereas 'no' suggests that the chance of infection only depends on whether or not the household is 'contaminated'. How to perform such a test is discussed in Becker (1989), section 2.6.2, where the Greenwood assumption is tested against the Reed–Frost assumption.

A drawback with the tests discussed in this section is that they assume homogeneous individuals. An important unsolved problem is to develop tests that allow for some heterogeneity between individuals, e.g. due to age and/or sex. Wrong conclusions might be drawn if we treat heterogeneous individuals as if they were homogeneous. For example, if many households contain two adults and two children, and children are more susceptible to the disease, then a true local spread might not be detected because most cases will be children and they are evenly spread among households.

### *3.2. The infectivity function*

Consider a community of uniformly mixing individuals. A function $B(u)$ giving a measure of how infectious a given infective individual is $u$ time units after becoming infected is called the infectivity function, or infectiousness function. This function is of great public health interest, but its estimation is hampered by not being able to observe which disease transmissions are caused by a given infective individual. Mainly its estimation has been restricted to the case where

$$B(u) = \begin{cases} \beta, & \text{if } X \leqslant u \leqslant X + Y, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where $X$ is the duration of the latent period and $Y$ is the duration of the infectious period.

In the general epidemic model it is assumed that $\Pr(X = 0) = 1$ and $Y$ has an exponential distribution, whereas in the multiparameter chain–binomial model it is assumed that both $X$ and $Y$ are essentially constant with $X$ large and $Y$ small. The statistical analyses described in Bailey (1975), chapter 15, and Becker (1989), chapter 4, assume various convenient parametric forms for the distributions of $X$ and $Y$, and illustrate the analyses with applications to data on household outbreaks.

It seems plausible that the true infectivity function is a continuous curve for each individual, with some variation possible between individuals. However, the lack of adequate data often makes this function unidentifiable. For example, the distribution of the final state is known only to depend on the distribution of $\int_0^\infty B(u)\,\mathrm{d}u$ and the latter class of distri-

butions is not larger than that with $B$ as in model (10). However, knowledge about the infectivity function is central when aiming to reduce disease transmission and practical advice on methods for the estimation of its shape and a description of the requisite data would be valuable.

A starting-point for the development of such methodology might be as follows. Suppose that the development of an infectious organism by each infected individual follows a similar time line, i.e. each infected individual has exactly the same infectivity function. Now consider data on households which include knowledge of the times of infection of the primary case and the next infection in the household. It seems feasible to estimate the common infectivity function nonparametrically from such data.

A start has been made to such analyses in partner studies associated with HIV and AIDS, where the time of infection could be ascertained because it was acquired by a blood transfusion and information about the time of infection of the sexual partner is obtained from a blood test, so his or her time of infection might be interval censored (Shiboski and Jewell, 1992).

## 4. Heterogeneity between individuals

Sources of heterogeneity can, loosely, be classified into heterogeneity due to differences in the characteristics of individuals and heterogeneity arising from community structure. In this section we are concerned with the first type of heterogeneity, postponing heterogeneity due to community structure to Section 5.

Individual heterogeneity may mean that individuals vary in their susceptibility to infection, or in how infectious the individual is when infected or both. Different levels of susceptibility are easier to estimate than different levels of infectivity; see Baker and Stevens (1995) and Britton (1998). A simple explanation for this is that a high susceptibility in a subgroup is reflected by a large proportion infected in that subgroup whereas high infectivity in a subgroup leads to a large proportion of those infected being infected by this subgroup, but details on who infects whom are rarely available. For this reason infectivity is often assumed identical and heterogeneity is considered to arise only from varying susceptibility, e.g. Rhodes *et al*. (1996). Since varying infectivity is equally important when aiming to control the spread of a disease, estimation procedures for different levels of infectivity, and determining what data are required for such an estimation, is an area requiring attention.

Factors that are responsible for heterogeneity between individuals can be either identifiable or unidentifiable. Examples of identifiable factors are age, sex, vaccination status and previous history of disease, whereas examples of unidentifiable factors are genetic variation, immunological variation, vaccination (if response to it is random) or any of the examples of the identifiable factors where they are not known by the investigator.

### *4.1. Identifiable heterogeneity between individuals*

The first natural question to pose is whether the spread of disease depends on certain specific characteristics of individuals. A quick and simple way to test whether individuals are heterogeneous in terms of susceptibility is to classify cases and non-cases according to their characteristics and to analyse the corresponding contingency table (Becker, 1989). Addy *et al*. (1991) tested various hypotheses concerning infection rates by using the ML ratio statistic, based on a stochastic epidemic model for a heterogeneous population, when analysing the outbreak of influenza A in Tecumseh, Michigan, among 567 households.

If individuals are known to be heterogeneous, or if some test of homogeneity has been rejected, the next step is estimation. If the population can be separated into a few homogeneous subpopulations we have what is called a *multitype epidemic*. The general epidemic of Section 2, for example, is easily generalized to this situation. In the homogeneous population an infectious individual has contact with any given susceptible individual at rate $\beta/n$. In a multitype population this contact rate becomes $\alpha_i\beta_j/n$, where $\alpha_i$ is the relative infectivity of the infectious individual and $\beta_j$ the relative susceptibility of the susceptible individual (this model assumption is known as proportionate mixing, e.g. Hethcote and Van Ark (1987), the most general case having arbitrary coefficients $\{\beta_{ij}\}$). Using techniques similar to those of Section 2.1, Britton (1998) showed that the relative susceptibilities are consistently estimated from final state data by

$$\hat{\beta}_i = \frac{1}{S_i(0)} + \frac{1}{S_i(0)-1} + \ldots + \frac{1}{S_i(\tau)+1} \approx -\log(1-\tilde{p}_i), \qquad i = 1, \ldots, k, \tag{11}$$

where $\tilde{p}_i = 1 - S_i(\tau)/S_i(0)$ is the observed proportion of $i$-individuals who became infected. For final state data the infectivities are unidentifiable, but if the epidemic is observed completely they can be estimated by using ML theory for counting processes. An interesting feature is that such estimators have a very slow rate of convergence if the corresponding susceptibility is identical with some other susceptibility (Britton, 1998). Models containing more parameters may be analysed when information about the actual contacts is also available. For example, Rhodes *et al*. (1996) used counting process theory to derive procedures to distinguish susceptibility from exposure to infection under this scenario.

The population cannot always be separated into a few homogeneous subgroups. For example, if age is assumed to have an effect on either susceptibility or infectivity then it would be appealing to treat age as a continuous covariate. So, instead of dividing the population into age groups (resulting in a multitype epidemic), it might be better to treat susceptibility or infectivity as a function with age as a dependent variable. Such a function could either be assumed to have some parametric form or it could be estimated nonparametrically. Estimation procedures for continuous explanatory variables are an area desiring further work in the future.

Heterogeneity arising from age differences deserves special mention since age changes over time. If studying an epidemic over a longer time period the age changes should therefore be accounted for. In this situation births, deaths, immigration and emigration, and perhaps waning immunity, should also be incorporated in the model. Some deterministic models with these features have been considered, but statistical analyses with such features require attention.

### 4.2. *Unidentifiable heterogeneity between individuals*

Inference for unidentifiable heterogeneities is only possible if data contain longitudinal information or if several epidemics are observed, e.g. many household outbreaks. A common way to model unobservable heterogeneities is by means of random effects, i.e. to model individual susceptibilities and/or infectivities as realizations of random variables, usually defined to be mutually independent. For household epidemics such models have been considered for example by Becker (1989), section 3, and Baker and Stevens (1995).

Becker and Yip (1989) treated the situation where one epidemic is observed completely and continuously by assuming variable susceptibility among individuals. It is shown that such heterogeneity typically results in the same type of disease progress as if the transmission parameter decreases over calendar time. The explanation appears to be that, over time,

it is the individuals with high resistance to the disease who tend to avoid infection, thus decreasing the rate at which new infections occur in the community.

Several parameter estimates have been derived from deterministic models containing heterogeneity, identifiable as well as unobservable, e.g. Anderson and May (1991) and Hethcote and Van Ark (1987). As mentioned previously, a drawback of obtaining estimators from deterministic models is that it does not provide an associated measure of their uncertainty.

There is always a risk that some sources of heterogeneity may be overlooked and an important question is what effect this has on the statistical analysis, in terms of understanding the dynamics of spread and when aiming to control the disease. Comparing the spread of disease in heterogeneous and homogeneous populations has received much attention in the probability literature, e.g. Ball (1985). In the statistical literature such comparisons have only recently begun, and mainly when considering heterogeneities due to community structure, e.g. Becker and Utev (1997). One way to assess the effects of different model assumptions from a statistical perspective is to compare the immunity coverage that is needed to prevent epidemics, as it would be estimated under the different model assumptions; see Section 5.

## 5. Epidemics in structured communities

### 5.1. Community structure and the prevention of epidemics

There has been considerable interest in comparing the outcomes of models for the transmission of disease in communities of homogeneous uniformly mixing individuals with outcomes for models which allow some heterogeneity. To make such comparisons meaningful it is necessary to calibrate the two communities in some way. For example, comparisons might be made subject to the initial average susceptibility of individuals being the same. This gives meaningful mathematical comparisons, but it is difficult to gain a useful practical insight from such a comparison. For practical insight it seems necessary to bring data into the comparison. For example, given data on one or more epidemics in a community we might then compare conclusions reached from these data under the assumption that the community consists of homogeneous uniformly mixing individuals with conclusions reached from an alternative analysis of the same data that allowed for some form of heterogeneity.

We now illustrate this point with respect to the following practical problem. Suppose that we wish to immunize a fraction of the members of a large community against a certain infectious disease with the aim of preventing future epidemics. The question is: how large must the immunity coverage be to prevent epidemics?

#### 5.1.1. Community of towns

In a uniformly mixing community, consisting entirely of susceptible individuals, the critical immunity coverage is $v_{\mathcal{U}} = 1 - 1/\theta_0$. This is true because the reproduction number after vaccinating a proportion $v$ is reduced to $(1 - v)\theta_0$, and when this number is less than or equal to 1 a major epidemic cannot occur (see Section 2.1). The basic reproduction number $\theta_0$ is estimated by expression (5) if data on just one large epidemic are available which give an estimate of the critical immunity coverage under the assumption of a uniformly mixing community.

In contrast, suppose that the community is made up of $k$ large localities, perhaps towns, with uniform mixing in each of these and negligible interaction between localities. Then locality $j$ has a basic reproduction number $\theta_{j0}$ and critical immunity coverage $v_j = 1 - 1/\theta_{j0}$, so the critical immunity coverage for the entire community is

$$v_{\mathcal{L}} = \sum_{j=1}^{k} \pi_j v_j,$$

where $\pi_j$ is the proportion of all individuals residing in locality $j$.

Becker and Utev (1997) showed that if we base estimation of the $\theta_{j0}$ on the same data set, where the large epidemic is comprised of a large epidemic in each of the localities, then the estimate of $v_{\mathcal{L}}$ is the same as the estimate of $v_{\mathcal{U}}$ only if all localities have the same basic reproduction number; otherwise $\hat{v}_{\mathcal{L}} > \hat{v}_{\mathcal{U}}$. In other words, the critical immunity coverage is underestimated if this kind of heterogeneity is present and we ignore it.

This comparison gives a useful practical insight. We now give another example.

#### *5.1.2. Community of households*

Suppose that the community consists of a large number of households and that individuals tend to mix more with members of their household than with other members of the community. The comparison is now complicated by the fact that the performance of a vaccination strategy depends on how immunizations are distributed over the households. One strategy is strategy H, which consists of independently selecting each household for complete immunization with probability $v_{\mathrm{H}}$. Households that are not selected remain completely susceptible under this strategy. The critical immunity coverage for this strategy will differ from that for strategy I, which immunizes each individual independently with probability $v_{\mathrm{I}}$.

Again it is of interest to compare the estimate of the critical immunity coverage under the assumption of a uniformly mixing community with that under the assumption of a household structure, when the same data are used. For a disease that is highly infectious within households it is found (Becker and Utev, 1997) that the estimate of the critical immunity coverage using strategy H is underestimated by assuming a uniformly mixing community. However, this is not necessarily true for other vaccination strategies, since some strategies perform better than strategy H. It is also of interest to know how these results depend on the distribution of household sizes. An investigation of this has recently been begun (Utev and Becker, 1997), but many open problems remain.

### *5.2. Parameter estimation using data from a community of households*

Early analyses of data on outbreaks in households were based on the assumption that, once the household has been infected, its outbreak evolves essentially independently of the presence of disease in the rest of the community. Epidemic chain models, as described in Section 2.2, formed the basis for these analyses. Although it is probably true that the probability of infection by a given infective household member is much larger than the probability of being infected by a given infective individual from outside the household, there are so many more infective individuals outside the household during a major epidemic that the probability of being infected from outside the household cannot be neglected. Some methods of parameter estimation that allow for disease transmission between households are now available.

Longini and Koopman (1982) proposed an analysis for data on a sample of households whose members had their blood tested for susceptibility before the epidemic season and again after the epidemic season. They adopted the Reed–Frost assumption for transmission between members of the same household. Transmission from other individuals is accommodated by introducing a global force of infection $\lambda(t)$, depending on time $t$, and supposing that each susceptible individual independently escapes infection from the global source, over the entire epidemic season $[0, T]$, with probability

$$q_B = \exp\left\{-\int_0^T \lambda(t)\,dt\right\}.$$

The quantity $q_B$ is treated as a parameter in their analysis. It actually depends on the size of the epidemic and therefore is more accurately viewed as a random variable. Nevertheless, treating $q_B$ as an unknown constant produces a relatively simple working model that enables an estimation of the probability of disease transmission between household members in a way that allows for the possibility of acquiring the disease from outside the household. ML estimation is used, with computation relying on the use of a set of recursive equations that specify the model probabilities.

Becker and Hopper (1983) were motivated by relatively complete data on an epidemic in a closed community. They used martingales for the underlying counting processes to derive explicit expressions for estimators of $\beta_W/\gamma$, the potential for infection within households, and $\beta_B/\gamma$, the potential for infection between households. Becker (1992) showed that the same methods, with essentially the same estimates, apply in the data setting considered by Longini and Koopman (1982), with the advantage of giving explicit expressions for estimates.

A more comprehensive method of analysis is described by Addy *et al.* (1991). They extended a model presented by Ball (1986). The model allows the infectious period to have any distribution, but computation becomes feasible only if its Laplace transform has an explicit expression. Individuals may acquire infection from outside the household by a global force of infection as in Longini and Koopman (1982). Discrete heterogeneity can also be accommodated, but in practice they reported computational difficulties when data on households with more than five members were included.

One statistical aspect that has been neglected is the design of studies associated with infectious diseases. To a large extent epidemic data arise from observing a disease that has run its natural course, but in some studies the statistical design of studies is relevant. An example occurs in the Tecumseh data analysed by Addy *et al.* (1991), where members of a sample of households had their sera tested before and after the epidemic season. Design problems associated with this type of study include the computation of the number of households required in the sample to enable estimation with a specified precision and what household sizes are best included in the sample. More specifically, are estimates more precise if we use 100 households with two initial susceptible individuals or 50 households with four initial susceptible individuals?

### 5.3. Geographic spread

The work by geographers on the spatial spread of infectious diseases is well illustrated in Cliff and Haggett (1993). They used both continuous time models and epidemic chain models, dividing the geography into a finite number of localities and then essentially treating individuals as different types, with type specifying their locality. It is then natural to allow individuals to change type, because of migration.

Time series models have also been used. They capture the dependence in the data only indirectly, because they do not propose to describe the transmission mechanism that generates the data.

Whereas there are deterministic models with a location included as a continuous variable (Bailey (1975), chapter 9), there does not seem to be any statistical analysis with a continuous location variable.

The collection of quality data on the transmission of disease is difficult under ideal

conditions. The study of the spatial spread of diseases requires data over a large region, adding enormously to the difficulty of gathering reliable and complete data, and thereby stifling the study of spatial spread. Although the concept of the velocity of an epidemic wave (Mollison, 1997; Metz and van den Bosch, 1995) is of great interest its formal estimation seems to be out of reach because of a lack of appropriate data.

## 6. Parameters crucial to the design of control measures

Sometimes we can side-step the difficulties of estimating all the parameters of an epidemic model by focusing on specific parameters of interest. We illustrate this with reference to estimating the fraction of the community that needs to be vaccinated to prevent epidemics and estimating the vaccine efficacy.

### *6.1. Estimating the critical immunity coverage*

When the general epidemic model applies to a large population the critical immunity level is $v_{\mathcal{U}} = 1 - 1/\theta_0$. Its estimation is manageable since the estimation of the basic reproduction number $\theta_0 = \beta/\gamma$ is manageable; see equation (5). We now consider the estimation of a certain critical immunity coverage for a community of households.

Let the community consist of a large number of households and suppose that immunization occurs by selecting households and vaccinating every member of each selected household. The critical immunity level under this vaccination strategy is $v_{\mathrm{H}} = 1 - 1/R_{\mathrm{H0}}$, where $R_{\mathrm{H0}}$ is the basic reproduction number for infected households. The mean number of households that an arbitrary infected household infects, if all other individuals were susceptible, is

$$R_{\mathrm{H0}} = \theta_{\mathrm{C0}} \nu_{\mathrm{H0}}$$

(Bartoszyński, 1972; Becker and Hall, 1996; Ball *et al.*, 1997), where $\theta_{\mathrm{C0}}$ is the mean number of individuals whom an infective individual would infect in households other than his own if everyone were susceptible and $\nu_{\mathrm{H0}}$ is the mean eventual number of cases in the household of an individual who is selected randomly from the community and infected, if every one of his household members is initially susceptible.

A direct estimation of $v_{\mathrm{H}}$ or $R_{\mathrm{H0}}$ does not seem feasible, and so we need to estimate $\theta_{\mathrm{C0}}$ and $\nu_{\mathrm{H0}}$. The former of these parameters is concerned with disease transmission between households; the latter with disease transmission within households. A full specification of the disease transmission model does not seem to be needed for the estimation of these parameters. More specifically, details of the distributions of latent and infectious periods are not needed; nor is a detailed specification of the way that disease spreads within the household. However, there is a serious challenge because the infection process is only partially observed. Both $\theta_{\mathrm{C0}}$ and $\nu_{\mathrm{H0}}$ are means and so they could be estimated by appropriate sample means if we could observe who infects whom for an epidemic in a community with all members initially susceptible. There are two problems with this. First, we cannot observe who infects whom. Second, the available data are usually from epidemics in a community in which a significant proportion of members are immune, as a result of either previous exposure to the disease or vaccination. Methods of estimation need to be devised that are based on observable parts of epidemics in a partially immune community.

An estimation of $\theta_{\mathrm{C0}}$ is possible from data on which households become infected and which avoid infection; see Becker (1995), section 2.2.2.

For the estimation of $\nu_{\mathrm{H0}}$ it helps to note that $\nu_{\mathrm{H0}} = \Sigma_n \, g_n \nu_n$, where $g_n$ is the proportion of

individuals who belong to households of size $n$ and $\nu_n$ is the mean size of the outbreak in a household where one of the $n$ initial susceptible individuals is infected. Census data on household sizes determine the $g_n$. Several possibilities exist for the estimation of the $\nu_n$. In the unlikely event that we have data on a number of household outbreaks for each household size $n$, where each household member was initially susceptible, we can estimate each $\nu_n$ by the corresponding sample mean $\bar{\nu}_n$. In the case of partial immunity we might be willing to assume that the size of the outbreak depends on the initial number of susceptible individuals, but not on the additional number of immune members of the household. Then we can manage with the sample mean of the size of household outbreak for each initial number of susceptible individuals. A problem is that with partial immunity we are unlikely to have many outbreaks in households with a large number of susceptible individuals, leading to imprecision in the estimate. When the total number of observed household outbreaks is moderate there seems no option other than to formulate a parametric model for the transmission of disease within households, which enables the use of data on all observed outbreaks to estimate the small number of parameters.

This is one example of how focusing on a specific objective can make parameter estimation feasible. We now give another such example.

### 6.2. Vaccine efficacy

Traditionally, vaccine efficacy has been defined as

$$\mathrm{VE} = 1 - \frac{\text{attack rate among vaccinated individuals}}{\text{attack rate among unvaccinated individuals}}$$

and this measure retains a central role in epidemiology circles. The attack rate is the proportion of individuals of that cohort who are infected over a specified time period. It is not a very satisfactory single measure of vaccine efficacy, because it depends on both the community from which the data come and on the time period over which the data are collected. Motivated by Smith *et al.* (1984), Halloran *et al.* (1992) made a more careful study of the interpretation and estimation of vaccine efficacy.

An interpretation of protective vaccine efficacy depends on the type of response that individuals have to the vaccine. It may be that, when a fully susceptible individual has a force of infection $\lambda(t)$ exerted on them at time $t$, then every vaccinated individual has a force of infection $\pi\lambda(t)$ exerted on them, where $\pi \in [0, 1]$ is a measure of the protection that the vaccine offers. However, it might be that a fraction $\pi$ of the vaccinated individuals are fully protected whereas the remainder have no protection at all. In each case $\pi$ can be considered a measure of protective vaccine efficacy and its estimation is of interest, but the interpretation is clearly different in the two cases.

It is likely that the responses to the vaccine vary between individuals and it would be interesting to devise studies that can test for heterogeneity in the response, perhaps by looking for overdispersion in the number infected in each of a number of groups.

## 7. Methodology for human immunodeficiency virus and acquired immune deficiency syndrome

There are very many references on statistical methodology for the study of the HIV epidemic. We review briefly a couple of major contributions and reflect on how these methods relate to, or may contribute to, the analysis of data on other infectious diseases.

### *7.1. Estimation of the incubation distribution*

The time that it takes from infection with HIV until diagnosis with AIDS is often called the incubation period. It is now known to have a median time of about 10 years, and to be highly variable. Its probability distribution is of interest for advising patients and as a tool for reconstructing the HIV incidence curve from observed AIDS incidence data; see Section 7.2. The estimation of this distribution presents difficulties because the time of infection is usually unknown.

There is one major group of infected subjects for whom the time of infection is known, namely those infected by receiving a transfusion of infected blood on a single occasion. We become aware of such subjects when they are diagnosed with AIDS, at which time their time of infection can be retrospectively ascertained. An estimation of the incubation distribution from such data requires care (Kalbfleisch and Lawless, 1989), since shorter periods are over-represented in such retrospectively ascertained incubation periods, because infected subjects with longer incubation periods have not developed AIDS by the time of the analysis.

The incubation distribution has also been estimated from survival times, without AIDS, observed in large cohort studies. The exact time of infection is generally not known, but regular blood tests for cohort members produce a time of last negative test to HIV antibodies and a time of first positive response. In other words, the time of infection is interval censored. Also, the time of AIDS-free survival is right censored. This has generated the development of extensions to Turnbull's (1976) self-consistency algorithm to cope with survival data that are doubly censored and truncated (DeGruttola and Lagakos, 1989; Sun, 1995).

The incubation periods of common diseases are considered reasonably well established (Benenson, 1990), but much of this knowledge is based on the accumulation of little bits of information gathered over a long time period, rather than on data from carefully designed studies. Therefore there is merit in studies and methods that formally estimate the distribution of the incubation period for the common diseases. The methods developed for the estimation of the incubation period of AIDS are relevant to other infectious diseases, but their application is impractical for most of them because the incubation periods are usually much shorter. These methods do, however, have potential for application to diseases, such as hepatitis, with a long latent period.

### *7.2. The method of back-projection*

Back-projection is a method for reconstructing the realized, but unobserved, HIV infection curve from the AIDS data by using knowledge about the incubation distribution. Its purpose, in the HIV and AIDS context, is to assess the extent of the HIV epidemic and to use the reconstruction as a basis for predicting AIDS incidences. This purpose has less relevance for most other infectious diseases, because epidemics of most diseases tend to be over before the data are available in a comprehensive form.

However, there is considerable interest in reconstructing the infection process for other diseases for the alternative purpose of taking advantage of the explicit expressions that are available for ML estimates of parameters when the process is fully observed. If we can use the available data to make a plausible reconstruction of the realized infection process, then we can substitute the reconstructed process into these explicit expressions. The hope is that the resulting estimates will perform well, and this warrants investigation because this is a way of making a very difficult estimation problem feasible.

Suppose that the removal process is fully observed and we propose to reconstruct unobserved parts of the infection process with the purpose of using the available data together

with reconstructed parts of the process to take advantage of estimators such as expressions (2). The fact that infection occurs by the transmission of disease from person to person is ignored in the method of back-projection. Instead, just for the reconstruction of the unobserved parts of the data, we assume that infections occur according to a non-homogeneous Poisson process with intensity $\lambda_t$ at time $t$. When we allow the form of $\lambda_t$ to be sufficiently flexible it can reproduce the infection intensity of any transmission model. Let $N_t$ denote the number of infections occurring over $(0, t]$ in a closed community. Then

$$E(N_t) = \int_0^t \lambda_x \, \mathrm{d}x = \Lambda_t, \quad \text{say},$$

so that an estimate $\hat{\Lambda}_t$ of $\Lambda_t$ for all $t$ can be thought of as a reconstruction of the unobserved process $\{N_t\}$. Then $\{S_t, I_t\}$, the unobserved part of the process, is reconstructed by $S_t = S_0 - \hat{\Lambda}_t$ and $I_t = I_0 + \hat{\Lambda}_t - R_t$.

Irrespective of the underlying process of infection, when infectious periods are 'assigned' to infected individuals independently, the removal process has intensity

$$\mu_t = \int_0^t \lambda_{t-u} \, \mathrm{d}F_D(u), \tag{12}$$

where $D$ is the duration of the infectious period and $F_D$ is its distribution function. Equation (12) assumes that there is no latent period. In discrete time, e.g. when dealing with daily data, $\lambda_t$ and $\mu_t$ represent the mean infection incidence and mean removal incidence respectively at time $t$. Estimates of $\lambda_1, \lambda_2, \ldots$ are then taken to be reconstructions of the number of individuals infected on days $1, 2, \ldots$ respectively.

Equation (12) is the basis of the method of back-projection. Typically we assume that $\{R_t\}$ is an observed Poisson process with intensity function $\mu_t$ and $F_D(u)$ is assumed known from past studies. Starting with these assumptions several approaches have been used to obtain an estimate of $\lambda_t$. Isham (1989) assumed that $\mu_t$ has a parametric form, estimated the parameters from the removal times and, using $\hat{\mu}_t$, deconvoluted equation (12) to obtain an estimate $\hat{\lambda}_t$. Day *et al.* (1989) and Taylor (1989) used a parametric form for $\lambda_t$ and estimated its parameters from the removal data using equation (12) in the likelihood. It has become popular to work in discrete time, leaving the $\lambda_t$ as separate parameters to be estimated, in the spirit of functional estimation. This is an ill-posed inverse problem (O'Sullivan, 1986) and smoothing needs to be imposed to stabilize the estimate. Becker *et al.* (1991) achieved this by adding a smoothing step to the EM algorithm for obtaining the ML estimates of the $\lambda_t$, whereas Bacchetti *et al.* (1993) used the penalized likelihood for obtaining a smooth estimate for the curve defined by the $\lambda_t$.

One attempt at the use of reconstructed infection curves in the estimates (2) has been made by Becker and Hasofer (1998), with promising results. Their approach is first to use kernel smoothing on the observed removal process to obtain $R_t^*$, a smoothed version of $R_t$, and then, for the general epidemic model, to use

$$I_t^* = \frac{1}{\gamma} \frac{\mathrm{d}R_t^*}{\mathrm{d}t}$$

as the reconstruction of $I_t$ with $\gamma$ assumed known.

The proposal to reconstruct the infection process and to use the 'complete' data set for parameter estimation is in the spirit of data augmentation methods (Tanner, 1996) that have proved useful for solving estimation problems that are difficult because of missing data,

censoring and partial observation. This immediately suggests that other data augmentation methods, such as Bayesian analyses using Markov chain Monte Carlo methods, have a role to play in the analysis of infectious disease data. A start has been made by Gibson and Renshaw (1998) and O'Neill and Roberts (1999) on ways to implement these methods for the analysis of epidemic data. An attractive aspect of such analyses is that it makes it possible, in principle, to allow for the heterogeneity that undoubtedly exists between individuals and households. In particular, in the analysis of data on household outbreaks it has been found necessary to allow for heterogeneity; see Bailey (1975), Becker (1989) and Baker and Stevens (1995). In analyses by ML methods this has required simplifying assumptions. The Markov chain Monte Carlo methods have the potential to provide analyses under realistic assumptions.

## 8. Important future directions

Several statistical problems have been mentioned that require attention. It seems useful to express an opinion about the statistical work, in the various areas, that is most worthwhile in terms of public health impact.

A literature search on vaccine trials and efficacy produces a vast number of papers describing different studies. This is not surprising since vaccination is often the most effective means of preventing the transmission of disease. Modellers and data analysts have much to contribute to this work, although they have only recently started to take up the challenge as a literature search on modelling work concerned with vaccine efficacy reveals. Real vaccines have shortcomings. Modellers could help substantially by quantifying the way that randomness in the response to a vaccine affects the effect that vaccination has on preventing the transmission of disease. Statisticians should tackle the associated estimation problems. More specifically, it is important to estimate the parameters that specify the vaccination coverage which is required to prevent epidemics, under various vaccination strategies when those vaccinated vary randomly in their response to vaccination.

The bulk of data collected on epidemics consists of surveillance data. These data tend to be of poor quality, usually suffering from severe underreporting and rates of reporting that vary over time. For example, some practitioners are inclined to pay more attention to the task of notifying the incidence of disease at times when many cases are occurring. The use of such unreliable data is clearly of doubtful value. Rather than trying to encourage all sources to report all cases, and obtaining data suffering from underreporting and variability in reporting, it may be more practical to obtain complete data from just a carefully chosen sample of the sources. Statisticians can help by providing proper advice on what type and how much data are required to make decisions on disease control policy that are objective and supported by reliable data.

A very rewarding consequence of developing new statistical methodology is when it motivates someone to conduct a study that specifically takes advantage of this methodology. This seems a possibility for the important problem of estimating the infectivity function for diseases. Hence the development of methods for such estimation and determining their data requirements are research topics of high priority.

In modelling work aimed at assessing the consequences of intervention through vaccination it is often acknowledged that disease transmission occurs at different rates in different age groups (Anderson and May (1991), chapter 9). However, the methods that have been used to estimate age-specific transmission rates assume that no immunization programme is in place and that the transmission process is in a steady state. For most diseases neither is true today and we urgently need methods for estimating age-specific transmission rates from data

on disease transmission in communities that contain some immunized individuals and avoiding the steady state assumption.

## Acknowledgements

This paper was written while Tom Britton spent a year at La Trobe University on a post-doctoral fellowship. He is grateful to the Swedish Natural Science Research Council for funding this visit. Niels Becker gratefully acknowledges support from the Australian Research Council.

## References

Addy, C. L., Longini, I. M. and Haber, M. (1991) A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, **47**, 961–974.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical Models based on Counting Processes*. New York: Springer.

Anderson, R. M. and May, R. M. (1991) *Infectious Diseases of Humans; Dynamic and Control*. Oxford: Oxford University Press.

Bacchetti, P., Segal, M. R. and Jewell, N. P. (1993) Backcalculation of HIV infection rates. *Statist. Sci.*, **8**, 82–119.

Bailey, N. T. J. (1975) *The Mathematical Theory of Infectious Diseases and Its Applications*. London: Griffin.

Baker, R. D. and Stevens, R. H. (1995) A random-effects model for analysis of infectious disease final-state data. *Biometrics*, **51**, 956–968.

Ball, F. G. (1983) The threshold behaviour of epidemic models. *J. Appl. Probab.*, **20**, 227–241.

———(1985) Deterministic and stochastic epidemics with several kinds of susceptibles. *Adv. Appl. Probab.*, **17**, 1–22.

———(1986) A unified approach to the distribution of total size and total area under the trajectory of the infectives in epidemic models. *Adv. Appl. Probab.*, **18**, 289–310.

Ball, F. G., Mollison, D. and Scalia-Tomba, G. (1997) Epidemics with two levels of mixing. *Ann. Appl. Probab.*, **7**, 46–89.

Bartoszyński, R. (1972) On a certain model of an epidemic. *Applic. Math.*, **13**, 139–151.

Becker, N. G. (1989) *Analysis of Infectious Disease Data*. London: Chapman and Hall.

———(1992) Analysis of infectious disease data from a sample of households. *Inst. Math. Stud. Lect. Notes*, **18**, 27–40.

———(1995) Estimation of parameters relevant for vaccination strategies. *Bull. Int. Statist. Inst.*, **56**, book 2, 1279–1289.

———(1997) Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. *Statist. Meth. Med. Res.*, **6**, 24–37.

Becker, N. G. and Hall, R. (1996) Immunisation levels for preventing epidemics in a community of households made up of individuals of different types. *Math. Biosci.*, **132**, 205–216.

Becker, N. G. and Hasofer, A. M. (1997) Estimation in epidemics with incomplete observations. *J. R. Statist. Soc.* B, **59**, 415–429.

———(1998) Estimating the transmission rate for a highly infectious disease. *Biometrics*, **54**, 730–738.

Becker, N. G. and Hopper, J. L. (1983) The infectiousness of a disease in a community of households. *Biometrika*, **70**, 29–39.

Becker, N. G. and Utev, S. (1997) The effect of community structure on the immunity coverage required to prevent epidemics. *Math. Biosci.*, **147**, 23–39.

Becker, N. G., Watson, L. F. and Carlin, J. B. (1991) A method of non-parametric back-projection and its application to AIDS data. *Statist. Med.*, **10**, 1527–1542.

Becker, N. G. and Yip, P. (1989) Analysis of variation in an infection rate. *Aust. J. Statist.*, **31**, 42–52.

Benenson, A. S. (ed.) (1990) *Control of Communicable Diseases in Man*, 15th edn. Washington: American Public Health Association.

Britton, T. (1997a) Tests to detect clustering of infected individuals within families. *Biometrics*, **53**, 98–109.

———(1997b) A test of homogeneity versus a specified heterogeneity in an epidemic model. *Math. Biosci.*, **141**, 79–99.

———(1998) Estimation in multitype epidemics. *J. R. Statist Soc.* B, **60**, 663–679.

Brookmeyer, R., Quinn, T., Shepherd, M., Mehendale, S., Rodrigues, J. and Bollinger, R. (1995) The AIDS epidemic in India: a new method for estimating current human immunodeficiency virus (HIV) incidence rates. *Am. J. Epidem.*, **142**, 709–713.

Cliff, A. D. and Haggett, P. (1993) Statistical modelling of measles and influenza outbreaks. *Statist. Meth. Med. Res.*, **2**, 43–73.

Day, N. E., Gore, S. M., McGee, M. A. and South, M. (1989) Predictions of the AIDS epidemic in the U.K.: the use of the back projection method. *Phil. Trans. R. Soc. Lond.* B, **325**, 123–134.

DeGruttola, V. and Lagakos, S. W. (1989) Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1–11.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.
Gibson, G. J. and Renshaw, E. (1998) Estimating parameters in stochastic compartmental models using Markov Chain methods. *IMA J. Math. Appl. Med. Biol.*, **15**, 19–40.
Halloran, M. E., Haber, M. and Longini, I. M. (1992) Interpretation and estimation of vaccine efficacy under heterogeneity. *Am. J. Epidem.*, **136**, 328–343.
Hethcote, H. W. and Van Ark, J. W. (1987) Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation and immunization programs. *Math. Biosci.*, **84**, 85–118.
Isham, V. (1989) Estimation of incidence of HIV infection. *Phil. Trans. R. Soc. Lond.* B, **325**, 113–121.
Isham, V. and Medley, G. (eds) (1989) *Models for Infectious Human Diseases: Their Structure and Relation to Data*. Cambridge: Cambridge University Press.
Kalbfleisch, J. D. and Lawless, J. F. (1989) Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *J. Am. Statist. Ass.*, **84**, 360–372.
Longini, I. M. and Koopman, J. S. (1982) Household and community transmission parameters from final distributions of infections in households. *Biometrics*, **38**, 115–126.
McKendrick, A. G. (1926) Applications of mathematics to medical problems. *Proc. Edinb. Math. Soc.*, **44**, 98–130.
Metz, J. A. J. and van den Bosch, F. (1995) Velocities of epidemic spread. In *Epidemic Models: Their Structure and Relation to Data* (ed. D. Mollison). Cambridge: Cambridge University Press.
Mollison, D. (1977) Spatial contact models for ecological and epidemic spread (with discussion). *J. R. Statist. Soc.* B, **39**, 283–326.
O'Neill, P. D. and Roberts, G. O. (1999) Bayesian inference for partially observed stochastic epidemics. *J. R. Statist. Soc.* A, **162**, 121–129.
O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, **4**, 502–518.
Rhodes, P. H., Halloran, M. E. and Longini, Jr, I. M. (1996) Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility. *J. R. Statist. Soc.* B, **58**, 751–762.
Rida, W. N. (1991) Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *J. R. Statist. Soc.* B, **53**, 269–283.
Saidel, T., Sokal, D., Rice, J., Buzingo, T. and Hassing, S. (1996) Validation of a method to estimate age-specific human immunodeficiency virus (HIV) incidence rates in developing countries using population-based seroprevalence data. *Am. J. Epidem.*, **144**, 214–223.
Shiboski, S. C. and Jewell, N. P. (1992) Statistical analysis of the time dependence of HIV infectivity based on partner study data. *J. Am. Statist. Ass.*, **87**, 360–372.
Smith, P. G., Rodrigues, L. C. and Fine, P. E. M. (1984) Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *Int. J. Epidem.*, **13**, 87–93.
Sun, J. (1995) Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics*, **51**, 1096–1104.
Tanner, M. A. (1996) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer.
Taylor, J. M. G. (1989) Models for the HIV infection and AIDS epidemic in the United States. *Statist. Med.*, **8**, 45–58.
Turnbull, B. W. (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc.* B, **38**, 290–295.
Utev, S. and Becker, N. G. (1997) Distribution functionals arising in epidemic control. *Technical Report 97-4*. School of Mathematical and Statistical Sciences, La Trobe University, Bundoora.