Statistical Studies of Disease Transmission: An Overview

Niels G. Becker and Tom Britton

School of Statistical Science, La Trobe University

1. Introduction

Infectious disease data have two features that distinguish it 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 clearly an integral part of statistical work in this area. While deterministic models can serve as 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.

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.

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 make the paper self-contained for readers unfamiliar with epidemic models and to act as a foundation for the discussion in the rest of the paper. Later sections are directed at more specific objectives. Methodology that can help to determine the way 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 and assessing the consequences of ignoring heterogeneity. The particular heterogeneity arising from having a structured community is investigated in greater detail in Section 5. A major motivation for the study of epidemic models is the insight 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 HIV, the virus that leads to AIDS. The analysis of AIDS data has attracted a wide range of statisticians who have developed methodology for the study of this, somewhat unique, epidemic. In Section 7 we focus on some of this methodology with a view to seeing if we can utilize these methods 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 SIR models in continuous time

For many transmissible diseases all individual are initially susceptible. Upon 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 an 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 introduced, in its stochastic form, by Bartlett [10]. It is a Markovian, continuous-time epidemic 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 $S(t)+I(t)+R(t) = n$ holds for all t, because the community is closed. Let $N(t) = I(t) + R(t)$, 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. Of course, this is rarely the case in real life. For example, 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 history of the epidemic process up to t, i.e. $\mathcal{H}_t$ is the sigma algebra generated by $\{S(u), I(u), R(u), N(u); 0 \leq u \leq t\}$. Then the General Epidemic is defined by

$$\begin{aligned}
\Pr(dN(t) = 1,\ dR(t) = 0 \,|\, \mathcal{H}_t) &\simeq \beta\overline{S}(t)I(t)dt, \\
\Pr(dN(t) = 0,\ dR(t) = 1 \,|\, \mathcal{H}_t) &\simeq \gamma I(t)dt, \\
\Pr(dN(t) = 0,\ dR(t) = 0 \,|\, \mathcal{H}_t) &\simeq 1 - \beta\overline{S}(t)I(t)dt - - - \gamma I(t)dt,
\end{aligned} \tag{1}$$

where $\overline{S}(t) = S(t)/n$ is the proportion susceptible at t (later we use the same notation for other variables). After some finite (random) time τ there are no infectious individuals present in the community and the epidemic is over. The number of individuals who are still susceptible and the number that have been removed specifies the *final state* of the epidemic.

The General Epidemic has two parameters, namely β and γ. The parameter β is the rate with which an infectious individual has close contacts with other members of the community, so $\beta\overline{S}(t)I(t)$ is the aggregated rate at which infectious individuals have close contacts with susceptibles. In statistical studies it is natural to transform the other parameter to γ^{-1}, the mean duration of the infectious period; the model implicitly assumes that the infectious period is exponentially distributed with parameter

γ.

The general epidemic model has several features that are open to criticism. The assumption of homogeneous individuals is relaxed in Section 4, while 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. Some resulting statistical issues are discussed briefly in Section 3.2.

MLE under complete observation

Assume that a realization of the General Epidemic is observed completely and continuously up to the end of the epidemic, i.e. we observe all of $\mathcal{H}_\tau$. Using counting process theory (e.g. [2, page 402]) we may then write the log-partial-likelihood explicitly as

$$\ell(\beta,\gamma) = \int_0^\tau \left(\log\left[\beta\overline{S}(u)I(u)\right] dN(u) - \beta\overline{S}(u)I(u)du + \log\left[\gamma I(u)\right] dR(u) - \gamma I(u)du \right).$$

The ML-estimators are obtained by maximizing this expression with respect to β and γ. By differentiating with respect to the parameters it is easily verified that the solutions are given by

$$\widehat{\beta} = \frac{N(\tau)}{\int_0^\tau \overline{S}(u)I(u)du} \qquad \text{and} \qquad \widehat{\gamma^{-1}} = \frac{\int_0^\tau I(u)du}{R(\tau) - r_0}. \tag{2}$$

Note that the numerator of the right term is simply the sum of all infectious periods and the denominator is the number of such infectious periods, so $\widehat{\gamma^{-1}}$ is simply the sample mean of the infectious periods.

Rida [40] shows that, on the part of the sample space were a major epidemic occurs, the ML-estimators are consistent and asymptotically independent normal variates. The standard errors of the estimators are given by

$$\text{s.e.}(\widehat{\beta}) = \frac{\widehat{\beta}}{\sqrt{N(\tau)}} \qquad \text{and} \qquad \text{s.e.}(\widehat{\gamma^{-1}}) = \frac{\widehat{\gamma^{-1}}}{\sqrt{R(\tau) - r_0}}.$$

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 and ML-estimation is cumbersome, making other approaches to statistical inference more attractive.

Martingale methods under incomplete observation

It might be that only the final state of the epidemic is observed. That is, besides the initial state (s_0, i_0, r_0) the data consist of $S(\tau), I(\tau)$ and $R(\tau)$ which actually contains only one random quantity since $I(\tau) = 0$ and hence $S(\tau) + R(\tau) = n$. Estimation procedures for this type of data have been proposed by several authors, e.g. [12, 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 that 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 [5, Section 6.82] and more recently by Becker and Hasofer [17, 18].

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 [5, page 118] or rely on large population approximations. In this section we use martingale techniques together with the method of moments to get explicit parameter estimates. To this end we define two zero mean $\mathcal{H}_t$-martingales M_1 and M_2 from the counting processes defined in (1):

$$M_1(t) = N(t) - \int_0^t \beta \overline{S}(u) I(u) du \tag{3}$$

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

¿From martingale theory it follows that

$$M(t) = \int_0^t \frac{1}{\overline{S}(u-)} dM_1(u) - \frac{\beta}{\gamma} M_2(t) = \frac{n}{s_0} + \frac{n}{s_0 - 1} + \ldots + \frac{n}{S(t) + 1} - \frac{\beta}{\gamma} [R(t) - - - r_0]$$

is a zero mean martingale (the last equality is a consequence of (3) and (4)). Note that $M(\tau)$ depends only on the initial and final states of the epidemic. Let $\theta = \beta/\gamma$, a quantity often called the basic reproduction number and denoted R_0. Then, by equating $M(\tau)$ to its mean (=0) we can estimate θ using only the initial and final states of the epidemic:

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

where $\tilde{p} = 1 - S(\tau)/s_0$ is the observed proportion who became infected. The estimator on the right can also be obtained from the determinstic version of the general epidemic. Applying the martingale central limit theorem it can be shown [12] (see [40] for more details) that, in a large community and in case of a major epidemic, the estimator above is approximately normally distributed with mean θ and a standard deviation which may be estimated consistently by

$$\text{s.e.}(\widehat{\theta}) = \left(\frac{1}{s_0^2} + \frac{1}{(s_0 - 1)^2} + \ldots + \frac{1}{[S(\tau) + 1]^2} + \frac{\widehat{\theta}^2}{n} \left[\overline{R}(\tau) - \overline{r}_0 \right] \right)^{1/2} \Big/ \left[\overline{R}(\tau) - \overline{r}_0 \right].$$

We are only able to estimate θ, or a function thereof, if only the final state is observed, which should not come as a surprise since we observe only one random quantity, $S(\tau)$

say $[I(\tau) = 0$ and $R(\tau) = n - S(\tau)]$. However, if the removal process is observed continuously one can do more. Becker and Hasofer [17] apply martingale techniques to obtain another estimating equation, which enables both β and γ to be estimated. The process they consider is $\tilde{M}(t) = S(t)\,(1 + \theta/n)^{R(t)}$ which can be shown to be a martingale using (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 explicit estimates and confidence regions for the two parameters β and γ.

It was shown previously that statistical analysis is quite straightforward when the epidemic is fully observed. Thus, if it were possible to reconstruct the complete epidemic process from the incomplete available data we would be able to 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 subsection we analyse 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. For example 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, 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 infected individuals of a given generation. The events that different susceptibles escape infection are assumed to be independent. Let I_k denote the number of infectives 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. $S_{k+1} = S_k - I_{k+1}$.

An epidemic chain is associated with a finite sequence of positive integers followed by a 0 (the 0 is often omitted in notation). The initial numbers of infectives and susceptibles, (i_0, s_0), are assumed given and the chain specifies the course of the spread. For example $1 \to 2 \to 1 \to 0$ means that $i_0 = 1, i_1 = 2, i_2 = 1, i_3 = 0$. The probability of a specific chain is obtained by sequential use of (6), and is found to be

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

In (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, Reed-Frost: $q_i = q^i$, and Greenwood: $q_i = q$. In the Reed-Frost case an individual has to 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'.

The model above can be generalized in several ways. For example, the probability of escaping infection may depend on the type of individual we are considering (multitype population), the specific generation and/or the household size. Another generalization is to assume that $\mathbf{q}$ is random with some specified distribution, *i.i.d.* for different individuals, resulting in a random effects model. For a more detailed description of such models we refer the reader to [12].

2.2.2 Epidemic chain data

If the epidemic chain is observed at each generation the statistical analysis is in principle straightforward just as is the case for complete observation in continuous time. The same is true for the various generalizations discussed in the previous subsection although these will not be treated further. Assume 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 household 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 given by

$$\ell(\mathbf{q};\mathbf{n}) = c + \sum_{s_0, i_0, \ldots, i_k} n(s_0, i_0, \ldots, i_k) \log\left[p_{s_0 i_0}(\mathbf{i};\mathbf{q})\right], \qquad (8)$$

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 $\mathbf{q}$. For the chain-binomial model, ignoring restrictions on the parameters, this can be done by differentiating (8) and setting the partial derivatives equal to 0. Let e_i denote the number of times 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 (n.b. these data are contained in the observed $\mathbf{n}$). It is not hard to show that the ML-estimators are given by $\hat{q}_i = e_i/(e_i + c_i)$. The estimators are consistent and asymptotically independent normal variates with a standard deviation which may be consistently estimated by s.e.$(\hat{q}_i) = [\hat{q}_i \hat{p}_i/(e_i + c_i)]^{1/2}$.

Under the Greenwood assumption ($q_i = q$) estimation is even simpler. Let $e = \sum_k e_k$ and $c = \sum_k c_k$. Then $\hat{q} = e/(e+c)$. Again the estimator is consistent and asymptotically normal with standard deviation estimated by s.e.$(\hat{q}) = [\hat{q}\hat{p}/(e+c)]^{1/2}$. For the Reed-Frost assumption ($q_i = q^i$) estimation is somewhat more complicated, in particular if some households are large, say of size six or more. The reason for this is that the likelihood contains terms like $p_i^r = (1-q^i)^r$, and to differentiate or maximize the resulting expressions is time-consuming when i and r are moderately large. There are several ways to overcome this problem. One way is to persist and explicitly differentiate all terms to obtain and to numerically solve the likelihood equation $\ell'(q; \mathbf{n}) = 0$, a polynomial equation of high degree. The variance could then be estimated, using the observed Fisher-information, by $[-\ell''(\hat{q}; \mathbf{n})]^{-1}$. Another way is to maximize the likelihood numerically. Since q is typically close to 1, the boundary of the parameter space, these methods may fail to obtain a solution. A different approach is by means of the EM-algorithm where data is treated as coming from a larger imaginary experiment. The event not to escape infection when exposed to i infected individuals, an event with probability $1-q^i = \sum_{j=0}^{i-1} \binom{i}{j} q^j p^{i-j}$, is replaced by i different imaginary events occuring with respective probabilities $\binom{i}{j} q^j p^{i-j}$, $j = 0, \ldots, i-1$. The EM-algorithm is a convenient tool for this maximization problem, because if we pretend that counts for the imaginary events are observed then the ML-estimation for the corresponding log-likelihood is simple, without any difficulty arising from the boundaries on the parameter space. First we compute the expectation (the E-step) of $\ell(q^{(r)})$ with respect to the imaginary data given the true data, and then maximize (the M-step) this expected log-likelihood with respect q to obtain a new estimator $q^{(r+1)}$. These two steps are repeated until the increase in the log-likelihood is negligible; see [30, 15] 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. That is, for each household the data consist of $r = \sum_{k \geq 1} i_k$, 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) who, besides the introductory case(s) had r ultimately infected individuals, and $\mathbf{m} = \{m(r; i_0, s_0)\}$. For these data the log-likelihood is given by

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

where $P_{s_0 i_0}(r; \mathbf{q}) = \sum_{\mathbf{i}:|\mathbf{i}|=r} p_{s_0 i_0}(\mathbf{i}; \mathbf{q})$ and the latter are defined in (7). In principle it is straightforward to maximize the likelihood but now numerical difficulties may appear. Since $P_{s_0 i_0}(r; \mathbf{q})$ is a sum of terms so its logarithm has complicated derivatives.

One way to overcome this difficulty is again by use of the EM-algorithm. The trick here is to *pretend* that the epidemic chain is available and it is done as follows. Start by picking your favorite guess $\mathbf{q}^{(0)}$ 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 $\mathbf{m}$. It is easy to show that the result satisfies

$$E\left(\ell(\mathbf{q};\mathbf{N})|\mathbf{M}=\mathbf{m};\mathbf{q}^{(0)}\right)=\sum_{s_0,i_0,r}\frac{m(r;s_0,i_0)}{P_{s_0i_0}(r;\mathbf{q}^{(0)})}\sum_{|\mathbf{i}|=r}p_{s_0i_0}(\mathbf{i};\mathbf{q}^{(0)})\log\left[p_{s_0i_0}(\mathbf{i};\mathbf{q})\right]. \quad (9)$$

The M-step stipulates that this expression should be maximized with respect to $\mathbf{q}$ to obtain a new estimate $\mathbf{q}^{(1)}$. The E- and the M-step are sequentially repeated, updating the estimate each time, until the increase of the log-likelihood at the new estimate value is negligible. This method might not seem to be a simplification of the problem, but it actually is. Firstly, $\mathbf{q}$ appears only in $\log[p_{s_0i_0}(\mathbf{i};\mathbf{q})]$ which is simpler to differentiate than $\log[P_{s_0i_0}(r;\mathbf{q})]$, and secondly, and perhaps more importantly, only the last factor in (9), $\log[p_{s_0i_0}(\mathbf{i};\mathbf{q})]$, depends on $\mathbf{q}$ over which we need to maximize.

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 $Bi(h,\pi)$ distribution truncated by excluding the zero class is appropriate. If uninfected households are also observed i_0 is modelled simply as coming from a $Bi(h,\pi)$ distribution. The same idea has been proposed for the analysis of other epidemic models [36, 1]. The EM-algorithm is then extended to *pretending* that the i_0's are also observed and the maximization is performed with respect to the additional parameter π.

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 without it analysis based on exact methods is practically impossible. Fortunately, for a model explicitly allowing infections between households Ball et al. [9] have shown that, given a major outbreak, household outbreaks actually 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 the chain-binomial model is not appropriate, at least in the case with separate q_i parameters. Instead, one then needs to resort to recursive formulas defining the distribution of the final size (e.g. [1]). Besides allowing latent and infectious periods to be random, it may be necessary to allow individuals to be of different types, for example 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 in the paper we have assumed that the disease spreads through person-to-person contacts. This is not always true. Alternative means of spread are, for example, through a commonly shared water source or transmission via air-borne organisms. Statistical analysis of data from such diseases differs in that there are no strong dependencies arising from interaction between infectives and susceptibles, 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. We refer to person-to-person transmission as *local* spread and transmission from some external source as *global* 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 is 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 [25] and references therein. Most proposed 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, N_r is the random quantity indicating how many cases there are in household r, and $f(\cdot)$ and $g(\cdot)$ are functions depending on the household size. The factor $\binom{N_r}{2}$ has the appealing interpretation to count the number of infected *pairs* of individuals in a household. 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. In [24] a test is derived for the same hypothesis but now assuming the epidemic is observed continuously. Britton [26] constructs a test for global spread against a general local structure, not necessarily a household structure. For example, if the population consists of animals, domestic or wild, or plants the natural local structure could be geographical such as the local structure *nearest neighbors*. 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

$$T = \sum_{i,j} d_{i,j}C_iC_j - \tilde{p}\sum_i d_{i,\cdot}C_i,$$

where $d_{i,\cdot} = \sum_j d_{i,j}$, C_i indicates whether or not individual j was a case and $\tilde{p}$ is the observed over-all proportion infected. Again, the distribution of T is computed under the null hypothesis of global spread and conditioned on $\tilde{p}$.

Assume now that either the spread is known to be local, or that some test has been performed and shown to be significant. The next step is then to find out more about the spreading mechanism. If treating chain-binomial models this can be done by assuming some structure on the q_i, the probability of escaping infection when exposed to i infected individuals of a given generation. For example, treating the spread within households, one might pose the question: Does the risk of being infected increase with the number of infectious individuals. A 'yes' would indicate person-to-person transmission whereas a '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 [12] who explicitly tests the Greenwood assumption against Reed-Frost. Similar tests can be derived for models described in continuous time although, for some reason, the Greenwood assumption has received less attention in the continuous time setting.

A drawback with the tests discussed in this section is that they assume homogeneous inviduals. An important unsolved problem is to develop tests that allow for some heterogeneity among individuals, e.g. due to age and/or sex. Wrong conclusions might be drawn if treating 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 $\Lambda(u)$ giving a measure of how infectious a given infective 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. For the most part its estimation has been restricted to the case where

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

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, while 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 [5, Chapter 15] and [12, 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.

While the form (10) for the infectivity function is convenient, it seems plausible that the true infectivity function is a continuous curve for each individuals, with some variation possible between individuals. The lack of adequate data often make it necessary to adopt a simple form like that in (10), but knowledge about the infectivity function is so central to the control of disease transmission and the use of appropriate models in our investigations that 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 infectious organism by each infected individual follows a similar time line. That is, each infected individual has the same deterministic infectivity function $\beta(u)$, $u \geq 0$. Now consider data on households which includes knowledge of the times of infection of the primary case and the next infection in the household. It seems feasible to estimate $\beta(u)$ nonparametrically from such data.

A start has been made to such analyses in partner studies associated with HIV/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 blood test, so his/her time of infection might be interval censored, [41].

4. Heterogeneity Among Individuals

Heterogeneities can loosely be classified into two different categories, heterogeneity due to individual factors (i.e. personal characteristics) and heterogeneity caused by community structure (e.g. environmental factors). In this section we will be concerned with the first type of heterogeneity, postponing the heterogeneity due to community structure to Section 5.

Individual heterogeneities can affect either the susceptibility: the risk of being infected, and/or the infectivity: how infectious an infected individual is. Different levels of susceptibility are easier to estimate than different levels of infectivity [6, 27]. A simple explanation for this phenomenon is that a high susceptibility in a subgroup is reflected by a large proportion infected in that subgroup. High infectivity in a subgroup, on the other hand, typically leads to a large proportion of the whole population being infected, largely by individuals from this subgroup, but details on who infects whom are rarely available. For this reason infectivities are often assumed identical and heterogeneity is only considered in terms of susceptibility (e.g. [39]). Still, it is intuitively clear that information about varying infectivities 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 estimation, is therefore an area deserving more attention in the future.

Factors responsible for heterogeneity among individuals can be either identifiable or unidentifiable. Examples of identifiable factors are age, sex, vaccination status and previous history of disease, while 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. Below we treat these two types of heterogeneity separately as their statistical analyses differ.

4.1 Identifiable heterogeneities among individuals

The first natural question to pose is if the spread of disease depends on certain specific characteristics of individuals. A quick and simple way to test if individuals are heterogeneous in terms of susceptibility, is to classify cases and noncases according to their characterstics and to analyse the corresponding contingency table [12]. Addy et al. [1] test different hypotheses concerning infection rates using the MLR-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 sub-populations we have what is called a *multitype epidemic*. For example, the general epidemic of Section 2 is easily generalized to this situation. In the homogeneous population an infectious individual has contact with any given susceptible at rate β/n. In a multitype population this contact rate might become $\alpha_i\beta_j/n$, where α_i is the relative infectivity of the infectious individual and β_j the relative susceptibility of the susceptible (this model assumption is known as proportionate mixing, e.g. [33], the most general case having arbitrary coefficients $\{\beta_{ij}\}$). Britton [27] considers estimation procedures for this model. Using techniques similar to those of Section 2.1, it is shown that the relative susceptibilities are consistently estimated from final state data by

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

where $\tilde{p}_i$ is the observed proportion of type-i individuals infected. For final state data the infectivities are unidentifiable, but if the epidemic is observed completely they can be estimated using ML-theory for counting processes. An interesting feature is that these estimators have a very slow rate of convergence if the corresponding susceptibilities are identical [27]. If information about the actual contacts is also available estimation in more complex models is possible. Using counting processes, Rhodes et al. [39] 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 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 is 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 also waning immunity, should be incorporated into the model. Some deterministic models with these features have been considered, but statistical analyses with such features require attention.

4.2 Unidentifiable heterogeneity among individuals

Inference for unidentifiable heterogeneities is only possible if data contains longitudinal information, or if several epidemics are observed, for example many household outbreaks. A common way to model unobservable heterogeneities is by means of random effects, that is 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 [12, Section 3] and Baker and Stevens [6]. The fit to data of the extended model can then be compared with the fit of the model without random effects. Applying certain recursive formulae when computing the probability of final state outcomes, such procedures are treated in [6], for models containing both identifiable heterogeneities and random effects.

Becker and Yip [22] treat the situation where one epidemic is observed completely and continuously assuming variable susceptibility amongst 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 individuals with high resistance to the disease will tend to remain susceptible thus decreasing the rate of new infections.

Several parameter estimates have been derived from deterministic models containing heterogeneity, identifiable as well unobservable (e.g. [3, 33]). As mentioned, a drawback of obtaining estimators from determinstic 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 for the statistical analysis, in terms of understanding the spreading dynamics 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. [7]. In the statistical literature such comparisons have only recently begun, and mainly when considering heterogeneities due to community structure, e.g. [20]. One way to the measure effects of different model assumptions from a statistical perspective is to compare the immunity coverage needed to prevent epidemics, as it would be estimated under the different model assumptions, see Section 5. For heterogeneities arising from differences in the characteristics of individuals such comparisons are yet to be made.

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 disease transmission 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 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 we wish to immunize a fraction of the members of the 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 susceptibles, the critical immunity coverage is $v_{\mathcal{U}} = 1-1/\theta_0$, where the basic reproduction number θ_0 is estimated by (5) if data on just one large epidemic is available. This gives an estimate of the critical immunity coverage under the assumption of a uniformly-mixing community.

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

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

where π_j is the proportion of all individuals residing in locality j.

Becker and Utev [20] show that if we base estimation of the θ_{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 $\widehat{v}_{\mathcal{L}} > \widehat{v}_{\mathcal{U}}$. 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_{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_{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 the case of a disease that is highly infectious within households it is found, [20], that the estimate of the critical immunity coverage using Strategy H tends to be 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 size. An investigation of this has recently been begun, [47], 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 the probability of acquiring infection from outside the household is negligible compared with the probability of being infected by an infective household member. In other words, once the household is infected its outbreak evolves essentially independently of disease presence in the rest of the community. Epidemic chain models, as described in Section 2.2, formed the basis for these analyses. While 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 from outside the household, there are so many more infectives 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 the possibility of disease transmission between households are now available.

Longini and Koopman [36] propose an analysis for data on a sample of households whose members had their blood tested for susceptibility prior to the epidemic season and again after the epidemic season. They adopt 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 independently escapes infection from the global source, over the entire epidemic season $[0, T]$, with probability

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

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 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. Maximum likelihood estimation is used, with computation relying on the use of a set of recursive equations that specify the model probabilities.

Becker and Hopper [19] are motivated by relatively complete data on an epidemic in a closed community. They use martingales for the underlying counting processes to derive explicit expressions for estimators of $\beta_{\mathrm{W}}/\gamma$, the potential for infection within households, and $\beta_{\mathrm{B}}/\gamma$, the potential for infection between households. Becker [13] shows that the same methods, with essentially the same estimates, apply in the data setting considered by Longini and Koopman, with the advantage of giving explicit expressions for estimates.

A more comprehensive method of analysis is described by Addy *et al.* [1]. They extend a model presented by Ball [8]. 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 [36]. Discrete heterogeneity can also be accommodated, but in practice they report computational difficulties when data on households with more than five members are 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 run its natural course, but in some studies the statistical design of studies is relevant. An example occurs in the Tecumseh data analyzed by Addy *et al.* [1], 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 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 susceptibles or 50 households with four initial susceptibles?

5.3 Geographic spread

The work by geographers on the spatial spread of infectious diseases is well illustrated in the paper by Cliff and Haggett [28]. They use 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.

While there are deterministic models with a location included as a continuous variable [5, Chapter 9], there does not seem to any statistical analysis with a contiuous location variable.

The collection of quality data on disease transmission 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. While concepts like the velocity of an epidemic wave, [37], have great theoretical interest they seem to be out of reach in practical terms because of a lack of data.

6. Parameters that are Crucial to the Design of Control Measures

Epidemic models can provide helpful guidance for the control of disease transmission, particularly if we can estimate the parameters of the model. We have seen that parameter estimation for epidemic models can be difficult. Sometimes we can sidestep those difficulties by focusing on a specific objective that does not require all of the parameters to be estimated. We illustrate this with reference to estimation of the fraction of the community that needs to be vaccinated to prevent epidemics and estimation of 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 $v_{\mathcal{U}} = 1 - 1/\theta_0$, where θ_0 is the basic reproduction number. We have already seen that estimation of $\theta_0 = \beta/\gamma$, see (5), is more managable than estimation of either β or γ. There is considerable interest in estimation of the critical immunity level for a structured community, and Becker [14] has considered the estimation of v_{H}, the critical immunity coverage for Strategy H in a community of households. Let us review the main issues for such estimation.

Consider a community consisting of a large number of households. The critical immunity level under Strategy H is $v_{\mathrm{H}} = 1 - 1/R_{\mathrm{H0}}$, where R_{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

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

[11, 16, 9], where θ_{C0} is the mean number of individuals an infective would infect in households other than his own if everyone were susceptible and ν_{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.

Direct estimation of v_{H} or R_{H0} does not seem feasible, and so we need to estimate θ_{C0} and ν_{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 disease spreads within the household. However, there is a serious challenge because the infection process is only partially observed. Both θ_{C0} and ν_{H0} are means and so they could be estimated by appropriate sample means if we could observe who infects whom for an epidemic occurring 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 occurring 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.

Estimation of θ_{C0} is possible from data on which households get infected and which ones avoid infection, [14, Section 2.2.2].

For the estimation of ν_{H0} it helps to note that $\nu_{\mathrm{H0}} = \sum_n g_n \nu_n$, where g_n is the proportion of individuals who belong to households of size n and ν_n is the mean outbreak size in a household where one of the n initial susceptibles is infected. Census data on household sizes determines the g_n. A number of possibilities exist for the estimation of the ν_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 ν_n by the corresponding sample mean $\overline{\nu_n}$. In the case of partial immunity we might be prepared to assume that the outbreak size depends on the initial number of susceptibles, but not on the additional number of immune household members. Then we can manage with the sample mean of size of household outbreak for each initial number of susceptibles. A problem is that with partial immunity we are unlikely to have many outbreaks in households with a large number of susceptibles, leading to imprecision in the estimate. When the total number of observed household outbreaks is moderate there seems no option but 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

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

and this measure retains a central role in epidemiology circles. 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.* [42], Halloran *et al.* [32] 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 individuals have to the vaccine. It may be that when a fully susceptible individual has a force of infection $\lambda(t)$ exerted upon them at time t, then every vaccinated individual has a force of infection $\pi\lambda(t)$ exerted upon them, where $\pi \in [0, 1]$ is a measure of the protection that the vaccine offers. On the other hand, it might be that a fraction π of the vaccinated individuals are fully protected while the remainder have no protection at all. In each case π 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 HIV/AIDS

There is a very large literature on statistical methodology for study of the HIV epidemic, and it is not our intention to review all of this work here. Instead, we focus on a couple of major contributions and reflect on how these contributions relate to, or may contribute to, the analysis of data on other infectious diseases.

7.1 Estimation of the incubation distribution

The time 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 the purpose of 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. Estimation of the incubation distribution from such data requires care, [35], since shorter periods are overrepresented in such retrospectively ascertained incubation periods, because infected subjects with longer incubation periods have not developed AIDS by the time of 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. As well, the time of AIDS-free survival is right censored. This has generated the development of extensions to Turnbull's self-consistency algorithm [46] to cope with survival data that are doubly-censored and truncated, [31, 43].

While these methods are applicable to other infectious diseases, they have considerably less scope for application because much shorter incubation periods usually apply. With diseases such as hepatitis there is some scope for application. It is unlikely that there is great interest in studies that determine the distribution of the incubation period for the common diseases since their properties are considered reasonably well established, [23]. This belief may not be well founded since much of this knowledge is based on the accumulation of little bits of knowledge over a long time period, rather than on data from carefully designed studies.

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 using knowledge about the incubation distribution. Its purpose, in the HIV/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 available for maximum likelihood 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 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 (2). The fact that infection occurs by disease transmission 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 λ_t at time t. When we allow the form of λ_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

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

so that an estimate $\widehat{\Lambda}_t$ of Λ_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 $S_t = S_0 - \widehat{\Lambda}_t$ and $I_t = I_0 + \widehat{\Lambda}_t - R_t$.

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, λ_t and μ_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 backprojection. Typically we assume that $\{R_t\}$ is an observed Poisson process with intensity function μ_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 λ_t. Isham [34] assumes μ_t has a parametric form, estimates the parameters from the removal times and, using $\widehat{\mu}_t$, deconvolutes (12) to get an estimate $\widehat{\lambda}_t$. Day *et al.* [29] and Taylor [45] use a parametric form for λ_t and estimate its parameters from the removal data using (12) in the likelihood. It has become popular to work in discrete time, leaving the λ_t as separate parameters to be estimated, in the spirit of functional estimation. This is an ill-posed inverse problem [38] and smoothing needs to be imposed to stabilize the estimate. Becker *et al.* [21] achieve this by adding a smoothing step to the EM algorithm for obtaining the maximum likelihood estimates of the λ_t, whereas Bacchetti et al. [4] use the penalized likelihood for obtaining a smooth estimate for the curve defined by the λ_t.

One attempt at the use of reconstructed infection curves in the estimates (2) has been made by Becker and Hasofer [17], with promising results. Their approach is to first use kernel smoothing on the observed removal process to get R^*_t, a smoothed version of R_t, and then, for the general epidemic model, use

$$I^*_t = \frac{1}{\gamma}\frac{\mathrm{d}R^*_t}{\mathrm{d}t}$$

as the reconstruction of I_t with γ assumed known.

The proposal to reconstruct the infection process and use the 'complete' data set for parameter estimation is in the spirit of data augmentation methods [44] that have proved useful for solving estimation problems that are difficult because of missing data, censoring and partial observation.

8. Work to be Done

A number of statistical problems have been mentioned that require attention. It seems worthwhile to express an opinion about the importance of further work in the various areas, in terms of public health impact.

A literature search on vaccine trials and efficacy will produce a vast number of papers describing different studies. This is not surprising since these studies are about the control of diseases and because drug companies have an obvious stake in such work. Modellers and data analysts have much to contribute to this work, although they have been slow to take up the challenge. A literature search on modelling work concerned with vaccine efficacy is quite modest. Modellers should acknowledge that vaccines are far from perfect, with their variety of responses in vaccinees and their undesirable side effects, and help to clarify the benefits, or otherwise, of vaccination programs in the light of the shortcomings of vaccines. Statisticians should address the associated estimation problems. A related issue is estimating the vaccination coverage required to prevent epidemics, under various vaccination strategies.

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 failing to record data that are essential if objective decisions on disease control are to be made. It is up to modellers and statisticians to provide proper advice on how to improve the surveillance data. Sometimes only a small fraction of disease cases are reported, even when the disease is reportable, and the rate of reporting is likely to vary over time. For example, some practitioners might be inclined to pay more attention to the task of notifying disease incidence at times when many cases are occurring. The use of such unreliable data is clearly of doubtful value. Rather than trying to get 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.

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 can happen when the objective of the proposed study is of sufficient importance. Estimation of the infectivity function for some diseases seems sufficiently important to bring this about. Hence the development of methods to do this and determining what data are required to make it possible seems to make it an area of high priority.

In modelling work to assess the consequences of intervention through vaccination it is often acknowledged that disease transmission occurs at different rates in different age groups, [3, Chapter 9]. However the methods that have been used to estimate age-specific transmission rates assume that there is no immunisation program and that the transmission process is in 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 immunised individuals and avoiding the steady state assumption.

Acknowledgements

Niels Becker gratefully acknowledges support from the Australian Research Council and Tom Britton is grateful to The Swedish Natural Science Research Council for financial support.

References

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

[2] Andersen P. K., Borgan Ø, Gill R. D., Keiding N. (1993). *Statistical models based on counting processes*. New York: Springer.

[3] Anderson R. M. and May R. M. (1991). *Infectious diseases of humans; dynamic and control*. Oxford: Oxford University Press.

[4] Bacchetti P., Segal M. R. and Jewell N. P. (1993). Backcalculation of HIV infection Rates. *Statistical Science* **8**, 82-119.

[5] Bailey N. T. J. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. London: Griffin.

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

[7] Ball F. G. (1985). Deterministic and stochastic epidemics with several kinds of susceptibles. *Adv. Appl. Prob.,* **17**, 1-22.

[8] Ball F. G. (1986). A unified approach to the distribution of total size and total area under the trajectory of the infectives in epidemic models. *Adv. Appl. Prob.* **18**,289-310.

[9] Ball F. G., Mollison D. and Scalia-Tomba G. (1996). Epidemics with two levels of mixing. *Ann. Appl. Prob.*, in press.

[10] Bartlett M. S. (1949). Some evolutionary stochastic processes. *J. Roy. Statist. Soc. B,* **11**, 211-229.

[11] Bartoszyński R. On a certain model of an epidemic. *Applicationes Mathematicae* XIII(2):139-151 (1972).

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

[13] Becker N. G. (1992). Analysis of infectious disease data from a sample of households. *Institute of Mathematical Studies Lecture Notes* **18**, 27-40, IV Basawa and RL Taylor (eds).

[14] Becker N. G. (1995). Estimation of parameters relevant for vaccination strategies. *Bull. Int. Statist. Inst.* **56**, Book 2, pp 1279-1289.

[15] Becker N. G. (1997). Uses of the EM algorithm in the analysis of data on HIV/AIDS and other infectious diseases. *Statist. Meth. Med. Research* **6**, in press.

[16] 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.

[17] Becker N. G. and Hasofer A. M. (1996). Estimation in epidemics with incomplete observations. *J. Roy. Statist. Soc. B*, in press.

[18] Becker N. G. and Hasofer A. M. (1997). Estimating the transmission rate for a highly infectious disease. Submitted.

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

[20] Becker N. G. and Utev S. (1997). The effect of community structure on the immunity coverage required to prevent epidemics. Submitted.

[21] Becker N. G., Watson L. F. and Carlin J. B. (1991). A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine* **10**, 1527-1542.

[22] Becker N. G. and Yip P. (1989). Analysis of variation in an infection rate. *Austral. J. Statist.* **31**, 42-52.

[23] Benenson A. S. (editor) (1990). *Control of Communicable Diseases in Man*, 15^{th} edition. Washington: American Public Health Association.

[24] Britton T. (1996). A test to detect within-family infectivity when the whole epidemic process is observed. *Scand. J. Stat.*, in press.

[25] Britton T. (1997). Tests to detect clustering of infected individuals within families. *Biometrics* **53**, 50-61.

[26] Britton T. (1997). A test of homogeneity versus a specified heterogeneity in an epidemic model. *Math. Biosci.*, in press.

[27] Britton T. (1997). Estimation in multitype epidemics. In progress.

[28] Cliff A. D. and Haggett P. (1993). Statistical modelling of measles and influenza outbreaks. *Statist. Meth. Med. Research* **2**, 43-73.

[29] 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. Roy. Soc. London, Series B*, **325**, 123-134.

[30] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the *EM* algorithm (with Discussion) J. Roy. Statist. Soc. B, **39**, 1-38.

[31] DeGruttola V. and Lagakos S. W. (1989). Analysis of Doubly-Censored Survival Data, with Application to AIDS. *Biometrics*, **45**, 1-11.

[32] Halloran M. E., Haber M. and Longini I. M. (1992). Interpretation and estimation of vaccine efficacy under heterogeneity. *Amer. J. Epidemiol.* **136**, 328-343.

[33] 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.

[34] Isham V. (1989). Estimation of incidence of HIV infection. *Philosophical Transactions of the Royal Society of London, Series B*, **325**, 113-121.

[35] Kalbfleisch J. D. and Lawless J. F. (1989). Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS. *J. Amer. Statist. Assoc.*, **84**, 360-372.

[36] Longini I. M. and Koopman J. S. (1982). Household and community transmission parameters from final distributions of infections in households. *Biometrics* **38**, 115-126.

[37] Mollison D. (1977). Spatial contact models for ecological and epidemic spread (with Discussion). *J. Roy. Statist. Soc. B* **39**, 283-326.

[38] O'Sullivan F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science* **4**, 502-518.

[39] Rhodes P. H., Halloran M. E. and Longini I. M. (1996). Counting process models for infectious disease data: distinguishing exposure to infection from susceptibility. *J.R. Statist. Soc. B,* **58**, 751-762.

[40] Rida W. N. (1991). Asymptotic proporties of some estimators for the infection rate in the general stochastic epidemic. *J.R. Statist. Soc. B,* **53**, 269-283.

[41] Shiboski S. C. and Jewell N. P. (1992). Statistical analysis of of the time dependence of HIV infectivity based on partner study data. *J. Amer. Statist. Assoc.* **87**, 360-372.

[42] Smith P. G., Rodrigues L. C. and Fine P. E. M. (1984). Assessment of the protective efficacy of vaccines against common diseases usinmg case-control and cohort studies. *Int. J. Epidemiol.* **13**, 87-93.

[43] 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.

[44] Tanner M. A. (1996). *Tools for Statistical Inference: Methods for the exploration of posterior distributions and likelihood functions.* New York: Springer.

[45] Taylor J. M. G. (1989). Models for the HIV infection and AIDS epidemic in the United States. *Statistics in Medicine*, **8**, 45-58.

[46] Turnbull B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated Data. *J. Roy. Statist. Soc. B*, **38**, 290-295.

[47] Utev S. and Becker N. G. (1997). Distribution functionals arising in epidemic control. Submitted.