Statistics of Infectious Diseases

Niels G Becker - Bundoora, Australia

1. Introduction

Infectious diseases are still a major cause of morbidity and mortality. We hear much in the media about the AIDS epidemic, but it should also be remembered that diseases such as malaria, schistosomiasis, filariasis, hookworm disease, trachoma, leprosy and onchocerciasis still affect hundreds of millions of people today. Even less serious diseases such as measles, rubella, hepatitis and influenza remain of great concern because complications are quite common.

There remains much scope for statisticians to contribute to the understanding and control of infectious diseases, since much remains unknown about properties of the diseases and their spread. Analysis of infectious disease data presents several challenges to statisticians. The main difficulties stem from the lack of detail and the dependence in the data. Infectious disease data do not result from planned experiments, but arise from epidemics occurring in nature. This makes it difficult to accumulate precise data and explains why existing data sets are often lacking in detail. Another reason for the lack of detail is that the infection process is only partially observable. The time and source of infection are usually not determined. The dependence in the data arises from the fact that infected cases are the cause of further infected cases.

2. The art of modelling

The statistical analysis of data generally involves a family of models. The choice of this family must be such that the models are both appropriate for the data and suited to the objectives of the study. How do we choose a model? This is an area where subjective decisions are made, but there are useful guidelines available. We give a brief discussion of these guidelines in the context of infectious disease data and illustrate them with reference to applications in later sections.

It is desirable to adopt a model which reflects the mechanisms that generate the data. Such models have many advantages. Firstly, they provide reassurance to medical

collaborators and thereby increase the chance that our work is taken seriously. A more concrete advantage is that such models are usually formulated in terms of parameters which are epidemiologically meaningful. This makes the models well suited to the task of testing epidemiological hypotheses and estimating key parameters. Considerations of how the data are generated often lead to parametric models. Statistical inferences based on parametric models are generally more efficient than nonparametric methods, although this greater efficiency is only meaningful when a correct choice of parametric family has been made. This is more likely when the parametric family is arrived at from considerations of how the data are generated. This, in turn, provides confidence that there is a real gain in efficiency from the use of such parametric models.

It is particularly important to use models which describe how the data are generated when the objective is to predict the future progress of the epidemic. Only for such models is there any basis for the hope that the model will continue to describe the epidemic when we project the model into the future.

It is also essential to use models which describe how the data are generated when the objective is to assess proposed control programmes. In most sciences, the effects of changes are assessed by analysing results of repeated experiments. This is not possible for infectious diseases and so one overcomes this difficulty by constructing a model which describes the essential features of an epidemic in the community and then uses the model to predict the consequences of introducing specific changes. More specifically, the use of a model to evaluate a vaccination campaign, for example, is based on the hope that when we make a change to the model which reflects the proposed campaign, then the model will respond in a way as to adequately describe the essential features of an epidemic in the corresponding partially vaccinated community. Only for models which describe how the data are generated is there any basis for this hope.

In how much detail should we model the mechanism which is thought to generate the data? A rough rule is to adopt the simplest model which contains the main characteristics of the spread and adequately describes the available data. By including more details than can be supported by the data we run the risk of imposing our own biases on the analysis.

Many epidemic models have been formulated, to allow for the specific characteristics of different diseases and types of data sets available. Many of these are described in the books of Bailey (1975) and Becker (1989). Here we give just two applications; one is concerned with measles data and the other with AIDS data. The models used in these applications require only familiarity with the binomial and Poisson models, and so are a useful source of interesting exercises for elementary statistics classes.

3. Household data on measles

Consider data on outbreaks of measles in households of size three. By size three we mean that there are initially three susceptible individuals in the household. The household may contain, as well, any number of individuals who are immune due to previous exposure to measles or vaccination. More specifically, consider the data shown in Table 1, which have been extracted from Tables 14.8 and 15.8 of Bailey (1975). These data were collected in Cirencester, UK, and in Providence, Rhode Island. Known characteristics of measles, specifically the fact that the rash occurs about 14 days after

infection (with little variation), enabled the deduction that these outbreaks were initiated by a single introductory case. Suppose our concern is to compare the sizes of outbreaks in these two locations.

TABLE 1
Frequency data for measles in households of three.
Fitted frequencies based on independence of classification.

Size of Outbreak	Cirencester		Providence	
	Observed	Fitted	Observed	Fitted
1	6	6.1	34	33.9
2	11	5.5	25	30.5
3	43	48.4	275	269.6

An immediate reaction might be to recognise Table 1 as a contingency table and to make the comparison accordingly. The chi-squared statistic (with 2 degrees of freedom) is computed to be

$$\Sigma[(o_i - e_i)^2/e_i] = 7.27,$$

which is significant at the 0.05 level. What can we deduce from this? Well, it suggests that the size of outbreak distributions for Cirencester and Providence differ. By comparing the observed with the (estimated) expected frequencies one sees that the apparent difference stems primarily from relatively fewer outbreaks of size two in Providence. This comparison in the present distribution-free setting is not well suited to providing a plausible epidemiological explanation for the apparent difference.

We now discuss a parametric approach which attempts to model the mechanism which generates the data. Let the random variable C denote the number of eventual cases when one of the three susceptible individuals introduces measles to the households. It is assumed that the risk of infection from outside the household is negligible when compared with infection by an infective within the household. Let p be the probability that a given susceptible is infected by a given infective from the same household at some time during the latter's infectious period.

Consider a household of size three with one of the three being the introductory infective. Assuming that the two remaining susceptibles are exposed to the introductory infectives independently, we find $Pr(C = 1) = q^2$, where $q = 1-p$. The size of the outbreak will be 2 if one of the two susceptibles is infected by the introductory case, while the other escapes infection by both the introductory case and the secondary case. As either of the two susceptibles could be the one to escape infection we find $Pr(C = 2) = 2pq^2$, whence $Pr(C = 3) = 1 - Pr(C = 1) - Pr(C = 2) = p^2(1 + 2q)$. This provides a parametric model for the data of Table 1.

To use the parametric model as a basis for comparing the Cirencester and Providence data of Table 1, we must fit the model to the data from each of the two locations. Let q_1 and q_2 denote the parameter q for Cirencester and Providence

respectively. We estimate q_1 and q_2 by the method of maximum likelihood. The likelihood function corresponding to the data of Table 1 is

$$L(q_1,q_2) = (q_1^2)^6(2p_1q_1^2)^{11}\{p_1^2(1+q_1)\}^{43}(q_2^2)^{34}(2p_2q_2^2)^{25}\{p_2^2(1+2q_2)\}^{275},$$

which leads to maximum likelihood estimates $\hat{q}_1 = 0.347$ and $\hat{q}_2 = 0.272$. The associated standard errors are s.e.$(\hat{q}_1) = 0.042$ and s.e.$(\hat{q}_2) = 0.018$.

One can now compare the data in Table 1 by testing the hypothesis $H_0 : q_1 = q_2$, a hypothesis about parameters which are directly related to the infectiousness of the disease. Such a test can be based on the statistic

$$(\hat{q}_1 - \hat{q}_2)/[\text{s.e.}(\hat{q}_1 - \hat{q}_2)] = 1.65.$$

This is not significant, although it is a borderline decision with the .05 level of significance and alternative $H_1 : q_1 > q_2$.

When a parametric approach is used one needs to check that the model provides an adequate description of the data. Table 2 shows the results of fitting the parametric probability model to the data of Table 1. The chi-squared goodness of fit statistics give 0.47 and 6.85, for Cirencester and Providence respectively. This reveals that while the model is adequate for the Cirencester data it is judged inadequate, at the 0.01 level, for the Providence data. This points to a difference between the two data sets, as before. The current model-based approach suggests a way of finding a possible explanation for this difference. Apparently an essential characteristic is missing from our description of the Providence measles outbreaks.

TABLE 2

Frequency data for measles in households of three.
Fitted frequencies based on epidemic model.

Size of Outbreak	Probability of Outbreak	Cirencester		Providence	
		Observed	Fitted	Observed	Fitted
1	q^2	6	7.2	34	24.7
2	$2pq^2$	11	9.5	25	36.0
3	$p^2(1+2q)$	43	43.3	275	273.3

What are some plausible modifications which might lead to an adequate model for the Providence data? We decide this by considering which of the underlying model assumptions might be violated. A first thought might be that the parameter q might not be the same for each susceptible/infective pair. For example, the duration of the infectious period might vary considerably between infected individuals. This means the parameter q will depend on which infective the susceptible is exposed to. Models which incorporate such heterogeneity are described by Becker (1989, Sections 3.1-3.3). It turns out that a model including this characteristic is not adequate for the Providence data. Perhaps this is not unexpected, because we would expect the above model to be

inadequate for the Cirencester data if this were the reason for the lack of fit.

To explain the difference between the two locations we should really be looking for a characteristic of the community, rather than a characteristic of the disease itself. Perhaps the households vary more in Providence, possibly due to varying degrees of crowding or levels of hygiene. Models with variations between households are described by Bailey (1975, Section 14.4) and Becker (1989, Sections 3.4 and 3.5). It turns out that a model with this characteristic gives an adequate description of the Providence data of Table 1. We can conclude that a plausible explanation for the apparent difference in the size of outbreak distributions for measles in households of Cirencester and Providence is that there is greater variation among households in Providence. It is possible that another characteristic could be found which can also explain the difference.

To summarise, we have gained more insight by using a model which reflects the way the data are generated.

4. Moving from measles to AIDS

Measles is one example of a classic infectious disease. Books have been written about models and methods of analysis for data on such diseases. Now new methods are being developed for data arising from the AIDS epidemic. Why is it necessary to develop new methods? Why not just use the models and methods of analysis already available? The reason is that AIDS, or infection with HIV, has some different characteristics to those of previously studied infectious diseases. One such characteristic is the long duration of the incubation period, which is the time from infection with HIV to diagnosis with AIDS. The incubation period is now thought to be about ten years, and varies considerably from case to case. A consequence of this long incubation period is that individuals are changing their behaviour during the course of the epidemic and direct results of this are only seen much later. This makes it difficult to place reliance on any specific model assumptions.

5. Predicting the incidence of AIDS cases

To make the discussion more specific let us consider the data on monthly AIDS counts in Australia, which are shown in Table 3. The objective is to predict the future incidence of AIDS cases, for the purpose of planning future health needs.

5.1 *Alternative methods*

Without any additional knowledge one might simply plot the counts on a graph, fit a simple exponential, polynomial, or other mathematically convenient curve to the data points, and extrapolate the curve to arrive at predictions. This was indeed done in the early days of the epidemic; see Curran et al. (1985) and McEvoy and Tillett (1985). This approach makes no attempt to use knowledge about the way the data are generated. The curves are simply a way of smoothing the data and there is little basis for the hope that extrapolations of the fitted curves will be close to future data. A related concern is that one cannot estimate the prevalence of HIV infection with this approach.

TABLE 3
Australian AIDS incidence data

Months	Year								Total
	1982	1983	1984	1985	1986	1987	1988	1989	
January	0	0	0	11	15	30	39	53	
February	0	0	0	11	14	25	42	47	
March	0	0	1	6	13	31	27	32	
April	0	1	2	9	14	17	28	25	
May	0	0	0	19	18	46	34	39	
June	0	0	2	8	18	34	41	42	
July	0	0	3	9	16	24	49	38	
August	0	1	6	4	23	27	45	48	
September	0	1	7	11	22	37	41	48	
October	0	0	6	10	30	29	53	50	
November	0	1	5	9	25	44	58		
December	1	2	11	11	13	26	40		
Total	1	6	43	118	221	370	497	422	1678

It seems appropriate to adopt a model which takes account of our knowledge of the possible modes of transmission of HIV and our knowledge of the symptom-free period of the disease. See Isham (1988) for a review of transmission models which have been proposed to describe the dynamics of HIV infection. Some recent attempts at modelling the HIV epidemic have been extremely ambitious by including dozens of variables and parameters. It could well be that eventually something useful will come out of studying these elaborate models, but for the immediate problem of predicting future AIDS incidence their use seems rather limited. The reason for this is that the models typically involve quantities such as number of IV drug users, number of homosexual men, rate of needle sharing, and rate of change of sexual partners. Reliable information about these quantities is very limited at this time. In short, the transmission models tend to be overspecified relative to the available data.

The favoured method by which predictions are currently made is the method of back-projection or back-calculation. This method uses the fact that an AIDS case is the result of infection with HIV followed by an incubation period. The method of back-projection does not require us to keep track of different risk groups or different modes of transmission.

5.2 *The method of back-projection*

A time which is clearly before the introduction of the disease is chosen as the time origin. AIDS counts are usually reported on a monthly basis, as in Table 3. We therefore present our discussion in discrete time and refer to the time unit as a month. Let H_t denote the number of individuals infected with HIV in month t. The process $\{H_1,H_2,...\}$ is typically assumed to be a discrete non-homogeneous Poisson process.

The number of AIDS cases diagnosed in month t, is denoted by A_t, t = 1,2,...,T, where T is the month beyond which no reliable AIDS incidence data are available. Let f_d be the probability that the duration of the incubation period is d months, d = 0,1,2,...

In this notation we have the relationship

$$E[A_t|H_1,H_2,...,H_t] = \sum_{i=1}^{t} H_i f_{t-i}$$

between the monthly number of AIDS cases and the monthly number of infections. Taking the expectation of both sides gives

$$\mu_t = \sum_{i=1}^{t} \lambda_i f_{t-i}, \qquad (1)$$

where $\mu_t = E[A_t]$ and $\lambda_i = E[H_i]$.

The incubation period distribution is assumed known in the method of back-projection. In practice the distribution has been estimated from data on transfusion associated AIDS cases and from cohort studies of specific risk groups, such as haemophiliacs.

Consider now why the amount of detail in the model (1) is well suited to the task of making predictions. One reason is that one only needs the AIDS incidence data and knowledge of the incubation period distribution. The AIDS counts are clearly available. Also, much data has been collected on the incubation period, although most of these data provide information primarily about the left hand half of the incubation period distribution. Relatively little is known, at this time, about the values of parameters contained in more detailed transmission models. In short, the degree of detail in the model (1) seems appropriate for the data currently available. A second reason is that the quantities in equation (1) are precisely the quantities which are affected by intervention. Thus treatment offered to asymptomatic patients since 1987 will directly affect the incubation period distribution, i.e. the f's, while education programmes on safe sex and clean needle programmes for drug users will directly affect the λ's.

The statistical nature of the problem is that we have observations on random variables $A_1,A_2,...,A_T$, whose means are given by (1), in which the f's are assumed known. The aim is to estimate the λ_t. This is an example of an ill-posed inverse problem; see O'Sullivan (1986) for an overview of such problems. The estimates of the λ_t tend to be unstable unless some way of smoothing is introduced. In the present context it is common to introduce this smoothing with the use of a parametric form $\lambda_t = g(t;\theta)$ and then obtain an estimate of θ; see Brookmeyer and Gail (1988), Day et al. (1989) and Taylor (1989). An alternative is to fit a parametric form in μ_t to the AIDS incidence data, substituting the fitted $\hat{\mu}_t$ into (1) and solving for the λ_t, as Isham (1989) has done. A drawback of the latter approach is that the solution tends to include some negative λ_t.

Recently Becker et al. (1990) have left the λ_t in nonparametric form and have smoothed the estimates via a weighted moving average applied at each iteration of the EM algorithm used to maximise the likelihood. This is an application of a method proposed by Silverman et al. (1990) in a different context. In the same spirit, Rosenberg and Gail (1990) suggest weakly parametric approaches, one of which achieves

smoothing by the use of splines.

One advantage of the nonparametric method of back-projection with smoothing is that it lets the data speak for themselves in determining which configuration best explains the observed AIDS counts. Another reassuring feature of the method is its additive property. To explain this property suppose we wish to apply the method of back-projection to AIDS counts from the states of Victoria and New South Wales in Australia. If one parametric family of models for the λ_t is appropriate for Victoria and another for New South Wales, then a third parametric family will be appropriate for the combined data. This means that the search for an appropriate parametric family of models gets very tedious when we wish to apply the method to each of a number of risk groups and geographic regions, as well as various combinations of these. The non-parametric approach does not share this difficulty, because the sum of two or more non-homogeneous Poisson processes is again a non-homogeneous Poisson process.

Statisticians who, for some reason, prefer to use the method of back-projection with a parametric model can use the smoothed nonparametric estimates of the λ_t as a way of guiding them to a suitable parametric model.

5.3 *Making predictions*

How is the model formulation (1) used to make predictions of AIDS incidence? By estimating the λ_t we are essentially determining which configurations $\{\lambda_1,\lambda_2,...,\lambda_T\}$ could plausibly have generated the observed AIDS incidence data. Estimates of the λ_t indicate how many individuals were infected during the various months since the start of the epidemic. One can then use equation (1) to project these numbers forward to indicate when we might expect these HIV infected individuals to be diagnosed with AIDS. More specifically, the expected number of AIDS cases during month $T+\tau$ is given by

$$\mu_{T+\tau} = \sum_{i=1}^{T+\tau} \lambda_i f_{T+\tau-i}. \qquad (2)$$

Note that it is not enough to substitute the estimates $\{\hat{\lambda}_1,\hat{\lambda}_2,...,\hat{\lambda}_T\}$. We also need to specify values for $\lambda_{T+1},...,\lambda_{T+\tau}$. There are no data to estimate these values and so we need to substitute a variety of plausible values for them. Fortunately, short-term predictions are not sensitive to the values of $\lambda_{T+1},...,\lambda_{T+\tau}$, because of the long incubation period. Note also that the estimates for the λ_t, with t near T, are quite imprecise. Again the long incubation period ensures that short-term predictions are not sensitive to this imprecision. Our predictions are much more sensitive to lack of knowledge about the right hand tail of the incubation period distribution. The incubation period is primarily a property of the disease and so we can estimate the incubation period distribution using data from the USA, where the epidemic has a longer history.

5.4 *Application to Australian data*

Consider now an application of the nonparametric method of back-projection to the Australian data of Table 3. For the incubation period distribution we assume the Weibull distribution, as estimated by Brookmeyer and Goedert (1989), given by the distribution function

$$F_x = 1 - \exp(-0.0021x^{2.516}).$$

See Becker et al. (1990) for details of the method of estimating the λ_t via the smoothed EM approach. Figure 1 shows the estimated configuration $\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_T\}$ superimposed on the graph of the AIDS data of Table 3. The area under this configuration is about 10,000 and this is the estimated number of HIV infected individuals up to October 1989. [This estimate is actually too low because no allowance has been made for the fact that treatment became available in October 1987.] Our estimate of the infection intensity suggests that the infection intensity has peaked. However, it must be remembered that estimates of the λ_t are not precise for t near T (October 1989) and so the estimated near-zero infection intensity near T must not be taken too seriously. Figure 1 clearly depicts the fact that there are very many infected individuals who do not yet have clinical AIDS. The expected number of AIDS cases is still increasing, as can be seen from the two curves of predictions shown in Figure 1. The lower curve is the overoptimistic prediction under the assumption that there are no further infections after October 1989. The upper curve is based on the, seemingly pessimistic, assumption that there has been a continuing rate of infection of 100 cases per month since the beginning of 1986.

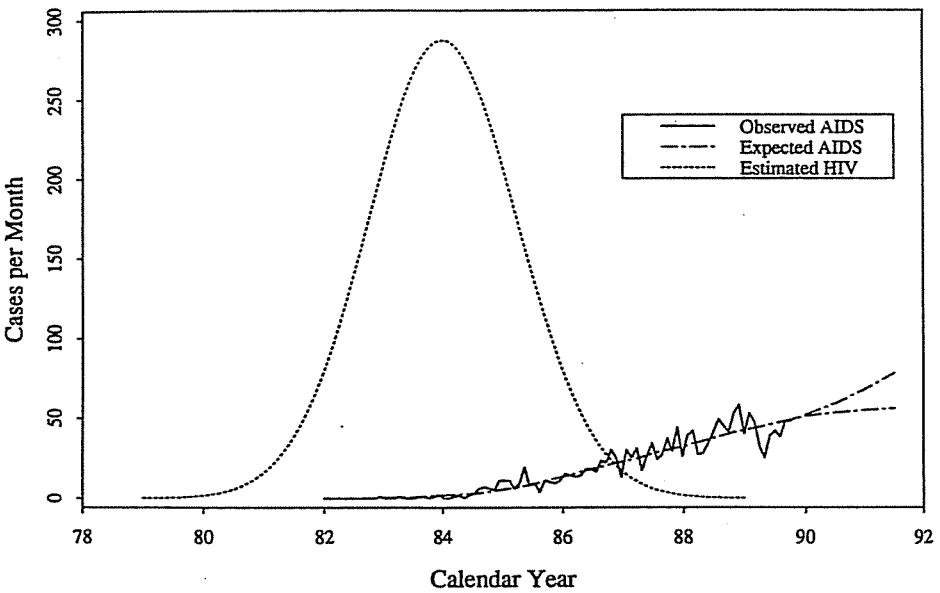

FIGURE 1
Observed monthly AIDS counts, estimated numbers of HIV infected individuals, and prediction of AIDS incidence

5.5 *Further considerations*

A number of considerations have not been discussed here. We mention just two very important ones. The first is that we need to indicate the precision of our

predictions. This can be done, and has been done, by simulation studies. The second is that the incubation period needs to be changed after 1987 to reflect the fact that there was a change in the definition of AIDS at that time and treatment with zidovudine (AZT) was introduced at about the same time. Solomon and Wilson (1990) and Brookmeyer and Liao (1990) have given two ways of incorporating the effect of treatment with AZT into the method of back-projection.

References

Bailey, N T J (1975) *The Mathematical Theory of Infectious Diseases and its Application.* Griffin, London.

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

Becker, N G, Watson, L W and Carlin, J B (1990) *A Method of Nonparametric Back-Projection and its Application to AIDS Data.* (Submitted for publication)

Brookmeyer, R and Gail, M H (1988) A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association* **83**, 301-308.

Brookmeyer, R and Goedert, J J (1989) Censoring in an epidemic with application to hemophiliac-associated AIDS. *Biometrics* **45**, 325-335.

Brookmeyer, R and Liao, J (1990) Statistical modelling of the AIDS epidemic for forcasting health care needs. *Biometrics* **46**. (To appear)

Curran, J W, Morgan, M W, Hardy, A M, Jaffe, H W, Darrow, W W and Dowdle, W R (1985) The epidemiology of AIDS : current status and future prospects. *Science* **229**, 1352-1357.

Day, N E, Gore, S M, McGee, M A and South, M (1989) Predictions of the AIDS epidemic in the UK : the use of the back projection method. *Philosophical Transactions of the Royal Society of London B* **325**, 123-134.

Taylor, J M (1989) Models for the HIV infection and AIDS epidemic in the United States. *Statistics in Medicine* **8**, 45-48.

McEvoy, M and Tillett, H E (1985) Some problems in the prediction of future numbers of cases of the acquired immunodeficiency syndrome in the UK. *Lancet* **ii**, 541-542.

O'Sullivan, F (1986) A statistical perspective on ill-posed inverse problems. *Statistical Science* **1**, 502-527.

Rosenberg, P S and Gail, M H (1990) Backcalculation of flexible linear models of the HIV infection curve. *Applied Statistics* **39**. (To appear)

Silverman, B W, Jones, M C, Wilson, J D and Nychka, D W (1990) A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *Journal of the Royal Statistical Society B* **52**, 271-324.

Solomon, P J and Wilson, S R (1990) Accommodating change due to treatment in the method of back projection for estimating HIV infection incidence. *Biometrics* **46**. (To appear)