BIOMETRICS 48, 117–128
March 1992

A Discrete-Time Model for the Statistical Analysis of Infectious Disease Incidence Data

Alvin H. Rampey, Jr.,[1] Ira M. Longini, Jr.,[2] Michael Haber,[2] and Arnold S. Monto[3]

[1] Statistical and Mathematical Sciences,
Lilly Research Laboratories, a division of Eli Lilly and Company,
Lilly Corporate Center, Indianapolis, Indiana 46285, U.S.A.
[2] Division of Biostatistics, Department of Epidemiology and Biostatistics,
Emory University, Atlanta, Georgia 30322, U.S.A.
[3] Department of Epidemiology, University of Michigan, 109 Observatory St.,
Ann Arbor, Michigan 48109, U.S.A.

SUMMARY

A discrete-time model is devised for the per-time-unit distribution of infectious disease cases in a sample of households. Using the time at which an individual is identified (e.g., when illness symptoms appear) as a marker for being infected, the probabilities of becoming infected from the community or from a single infectious household member are estimated for various risk factor levels. Maximum likelihood procedures for estimating the model parameters are given. An individual may be classified with regard to level of susceptibility and level of infectiousness. The model is fitted to a combination of symptom and viral culture data from a rhinovirus epidemic in Tecumseh, Michigan. In general, it is observed that decreasing risk of infection is associated with increasing age.

1. Introduction

An initial step in the control of many infectious diseases is to determine how various individual and environmental risk factors contribute to the transmission of the infectious agent within a household or community. This information can be used to determine the appropriate distribution of control measures, e.g., vaccination, or modification of risk factors that are known to affect the spread of a particular infectious agent.

In this paper, we present an incidence data model developed by Rampey (unpublished Ph.D. dissertation, Emory University, 1988) that complements the final attack rate data model of Longini and Koopman (1982) by incorporating additional information regarding the time of an observed event that can be linked to infection (e.g., onset of illness) and hence, the order in which such events occur within a given household. The Longini–Koopman model uses the distribution of the total number of infections or illnesses in households from a homogeneous community. Their model provides estimates of separate parameters describing community and within-household disease transmission. It requires data consisting of a sample of households observed at the beginning and at the end of a specified period of time, usually corresponding to the course of an epidemic. At the end of this period, the observed number of newly infected individuals per household is determined. From this information, Longini and Koopman estimate (i) the probability that during the time period a susceptible individual escapes infection from the community and (ii) the probability that a susceptible person escapes being infected by a single infective within the same household.

Key words: Incidence data; Infection from the community; Infection within the household; Maximum likelihood; Rhinovirus.

Haber, Longini, and Cotsonis (1988) have extended the Longini–Koopman model to assess the impact of risk factors on household and community sources of infectious agent transmission. Their extension can be used to identify risk factors affecting transmission within the household and risk factors affecting transmission in the community. Haber et al. (1988) discuss the application of these models to household data,which include the values of risk factors on the household level, and to individual data, which include the values of individual risk factors. Longini et al. (1988) use the model to investigate the effect of pre-epidemic antibody level and age on influenza A(H3N2) transmission in Tecumseh, Michigan. Addy (unpublished Ph.D. dissertation, Emory University, 1988) and Addy, Longini, and Haber (1991) extend the Longini–Koopman model to the case where the infectious period of the agent in question follows any continuous probability distribution.

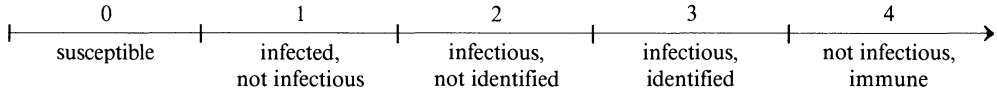
2. The Incidence Model

Assume that individuals in a household are observed starting at some time $t = 1$ and ending at some fixed time point $t = T$. Although several households may be under surveillance simultaneously, all times are measured in discrete units from the same time origin for individuals within a given household. The time origin does not have to be the same for all households; i.e., households may enter the study at different times. It is assumed that if an individual becomes infected, he or she will be identified as such either by observing the onset of symptoms and/or by some other means, such as analyzing throat or nasal cultures for the presence of virus. The time unit at which the individual is identified will then be recorded relative to $t = 1$ for the individual's household. Identification may be delayed until some time after the unobserved time of infection. Additional assumptions underlying the incidence model are as follows:

(i) A person may become infected any number of times during the course of an epidemic. However, a person reporting symptoms during two time units not separated by a minimum number of symptom-free time units will be assumed to have a single prolonged episode rather than a new infection.
(ii) Each individual belongs to a household containing one or more initially susceptible individuals.
(iii) Each person can be infected either from within the household or from the community.
(iv) The probability that a person is infected from the community is independent of the number of infected members in his or her household.

We also assume that the stopping time T is chosen sufficiently large so that the observation period $(1, T)$ covers the entire length of the epidemic with high probability. If the entire epidemic is not observed, then there will be individuals who become infected near the end of the observation period, but who do not develop identifying symptoms until afterward. Such individuals are incorrectly classified as having escaped infection. This may result in an underestimation of the disease transmission rates. However, this bias is assumed to be negligible if the observation period is long relative to the length of the epidemic.

All events (i.e., becoming infected, infectious, immune, or being identified) are assumed to occur at the beginning of a discrete time unit (e.g., a day or a week). A person can be in one of five states, coded as follows: 0—susceptible; 1—infected but not infectious, i.e., latent; 2—infectious but not yet identified; 3—infectious and identified, e.g., symptomatic; 4—immune and no longer infectious. An immune person may either return to state 0 after a certain time or stay in state 4 until the end of the study. The course of the infection for an individual who becomes infected could be represented as follows:



Let T_k ($k = 1, 2, 3$) be the discrete random variable that denotes the number of time units a person spends in state k. The T_k are nonnegative random variables with known probability mass functions. In classical epidemiological terminology, T_1 is the length of the latent period and $T_2 + T_3$ is the length of the infectious period. If individuals are identified by illness, then $T_1 + T_2$ is the length of the incubation period. If an immune person returns to the susceptibility state, then the number of time units he or she remains immune is assumed to be a known constant, t_4. For the same individual, T_1 and T_2 are assumed to be independent. The vectors $\mathbf{T} = (T_1, T_2, T_3)$ are independent and identically distributed over different individuals. Note that for two members of the same household, individuals i_1 and i_2, the events that individual i_1 is infectious at time unit t and that individual i_2 is infectious at time unit t are independent given the time units at which individuals i_1 and i_2 were identified.

Now, consider a given individual i. Define:

d_i = the time unit at which individual i became identified, or $d_i = \infty$ if individual i was never identified (d_i is the only observable information other than appropriate risk levels).

$f_i(t)$ = the probability that individual i was infectious at time unit t, given that individual i was identified at time unit d_i:

$$f_i(t) = \begin{cases} \Pr(T_2 \geq d_i - t), & \text{if } t < d_i, \\ \Pr(T_3 \geq t - d_i + 1), & \text{if } t \geq d_i. \end{cases}$$

$g_i(t)$ = the probability that individual i became infected at time unit t, given that individual i was identified at time unit d_i:

$$g_i(t) = \Pr(T_1 + T_2 = d_i - t).$$

Note that $g_i(t) = 0$ for $t > d_i$ and $g_i(t) = 0$ for all t when $d_i = \infty$.

We assume that the probabilities underlying the transmission of the infectious agent from one person to another depend on some characteristics of these individuals. More specifically, individuals are classified according to their susceptibility and infectiousness. For individual i, let r_i and h_i denote the levels of susceptibility and infectiousness, respectively. Define:

b_{r_i} = per-time-unit probability that individual i escapes infection from the community.

$q_{r_i h_j}$ = per-time-unit probability that individual i escapes infection from an infected household member j.

$e_i(t)$ = probability that individual i escaped infection at time unit t (if individual i is susceptible at that time unit), given only the d's of all other household members:

$$e_i(t) = b_{r_i} \prod_{j:j\neq i} \{f_j(t) q_{r_i h_j} + [1 - f_j(t)]\},$$

where the product is over all the other individuals in the same household.

Now, to define the likelihood function, consider first the case where an individual in state 4 must remain in this state until the end of the study. Let L_i denote the contribution of individual i to the likelihood function. For an individual who was never identified (i.e., $d_i = \infty$),

$$L_i = \prod_{t=1}^{T} e_i(t). \tag{2.1}$$

For an individual who was identified, the contribution to the likelihood function given that he or she became infected at time unit t is

$$L_i(t) = \prod_{u=1}^{t-1} e_i(u)[1 - e_i(t)] \tag{2.2}$$

and the total contribution to the likelihood function is

$$L_i = \sum_{t=1}^{T} L_i(t)g_i(t). \tag{2.3}$$

If the infected individual is to be returned to the pool of susceptibles after a fixed period of immunity, given by t_4, then additional contributions to the likelihood function are calculated. The entire observation period is broken into episodes in which the individual is known to be susceptible. For each of these episodes, L_i is calculated from (2.2) and (2.3) or from (2.1), depending on whether this individual was or was not identified again, respectively. For these additional contributions, the product taken over t in (2.1) starts at the time at which the individual is returned to the pool of susceptibles. The calculations in (2.2) and (2.3) correspond to the time period from the individual's return to the state of susceptibility until the last time unit at which he or she could become infected again, given the time unit of the next identification. The total contribution of individual i is the product of all his or her contributions L_i over all the episodes of susceptibility. The overall likelihood function is the product of the total contributions of all the individuals in all households.

The parameters of interest are the escape probabilities b_r and q_{rh} (defined above), which govern the process of transmission of the infectious agent. These probabilities may depend on one or more risk factors that affect the susceptibility and/or the infectiousness of the individuals. Maximum likelihood estimates of the parameters b_r and q_{rh} are found by numerically maximizing the likelihood function. Estimates of the covariance matrix of the parameter estimates are also obtained. The computer subroutines DB2ONF, DLINRG, DFDHES, and DLINDS in the IMSL library are used to find the maximum likelihood estimates and their variance–covariance matrix (IMSL, 1987). Starting values for the IMSL subroutines are determined by assuming T_1, T_2, and T_3 are fixed at their mean values and then applying a weighted least squares procedure (Rampey, unpublished dissertation cited previously). The parameter estimates thus obtained are adequate final estimates if T_1, T_2, and T_3 are fixed.

3. Hypothesis Testing

The general linear test approach is used to test hypotheses. For example, suppose we want to test the hypothesis that the community escape probabilities are the same for all the susceptibility levels:

$$H_0: \quad b_1 = b_2 = \cdots = b_r = \cdots.$$

This hypothesis can then be expressed in matrix notation as

$$H_0: \quad \mathbf{Cp} = \mathbf{0},$$

where $\mathbf{p}$ is the array of all the b_r and q_{rh} parameters and $\mathbf{C}$ is an appropriately defined matrix. Let $\hat{\mathbf{p}}$ denote the estimate of $\mathbf{p}$ and let $\mathbf{S}$ denote the estimated covariance matrix of $\hat{\mathbf{p}}$. Then

$$\chi^2 = (\mathbf{C}\hat{\mathbf{p}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\hat{\mathbf{p}})$$

is asymptotically distributed under the null hypothesis as a chi-squared random variable with k degrees of freedom, where k is the rank of the $\mathbf{C}$ matrix (Grizzle, Starmer, and Koch,

1969). Similarly, one could test the hypothesis that the household escape probabilities q_{rh} do not depend on the susceptibility level r or on the infectiousness level h or both. Generally, any null hypothesis that can be expressed in the matrix form $\mathbf{Cp} = \mathbf{0}$ can be tested using this procedure. It should be noted that if any of the parameters fall on the boundary of the parameter space then the corresponding columns of $\mathbf{C}$ and elements of $\mathbf{p}$ can be deleted, resulting in a hypothesis test that is conditional upon those parameters being at the boundary of the parameter space. Pairwise multiple comparisons can also be made using appropriate techniques to control the overall level of significance.

4. The Community Probabilities of Infection and Secondary Attack Rates

In epidemiological studies, one often wishes to calculate the community probability of infection (CPI) and the household secondary attack rate (SAR), as introduced by Longini et al. (1982). The CPI, used to measure the community involvement in disease transmission, is calculated as $1 - B_r$, where B_r denotes the probability that an individual at susceptibility level r escapes infection from the community over the entire course of the epidemic or over some other period of time. Thus, $B_r = b_r^T$, where T is the number of time units in the period of interest, and the community probability of infection is given by

$$\mathrm{CPI}_r = 1 - B_r = 1 - b_r^T. \tag{4.1}$$

The secondary attack rate (SAR) is used to measure the secondary transmission of infection within households. It is defined as the expected value of $100 \times (1 - Q_{rh})$, where Q_{rh} is the probability that a susceptible person at susceptibility level r escapes infection from a single infectious household member at infectiousness level h during the entire period when the infecting individual is infectious. Therefore, $Q_{rh} = (q_{rh})^{T_\mathrm{I}}$, where $T_\mathrm{I} = T_2 + T_3$ is the length of the infectious period. Hence, the household secondary attack rate for a susceptible individual at risk level r exposed to an infectious individual at infectiousness level h, is given by

$$\begin{aligned}\mathrm{SAR}_{rh} &= 100 \times \mathrm{E}\{1 - Q_{rh}\} = 100 \times [1 - \mathrm{E}\{(q_{rh})^{T_\mathrm{I}}\}] \\ &= 100 \times \left[1 - \sum_t q_{rh}^t \Pr(T_\mathrm{I} = t)\right].\end{aligned} \tag{4.2}$$

The SAR is important because it is a direct measure of how infectious a particular agent is given the type of exposure observed (see Longini et al., 1982, 1988).

An estimate of the variance–covariance matrix for the CPIs and SARs can be obtained using the method of statistical differentials and the expressions given above for the CPIs and SARs (Kendall and Stuart, 1977). Let

$$\mathbf{p} = (b_1, b_2, \ldots, q_{11}, q_{12}, \ldots)' \tag{4.3}$$

and

$$\mathbf{Y} = (\mathrm{CPI}_1, \mathrm{CPI}_2, \ldots, \mathrm{SAR}_{11}, \mathrm{SAR}_{12}, \ldots)', \tag{4.4}$$

where the CPIs and SARs are functions of the elements of $\mathbf{p}$ as defined above. Define

$$\mathbf{J} = \left[\left\{\frac{\partial y_i}{\partial p_j}\right\}\right]_{\mathbf{p}=\hat{\mathbf{p}}} \tag{4.5}$$

as the matrix of first partial derivatives of $\mathbf{Y}$ with respect to the elements of $\mathbf{p}$, evaluated at $\hat{\mathbf{p}}$. If $\mathbf{S}$ denotes the covariance matrix of $\mathbf{p}$, then

$$\mathrm{var}(\mathbf{Y}) \approx \mathbf{JSJ}'. \tag{4.6}$$

Now, hypothesis tests concerning the CPIs and SARs can be conducted using the general linear test approach described in Section 3.

5. The Rhinovirus Study in Tecumseh, Michigan

To illustrate the use of the incidence data model, we use data on rhinovirus transmission collected by Monto, Schwartz, and Albrecht (unpublished manuscript, 1988) in Tecumseh, Michigan, during the fall of 1983. Rhinovirus seasons in Tecumseh tend to run from early September to early November and generally have a duration of about 8 weeks (Longini, Monto, and Koopman, 1984). The virus is transmitted by direct contact with infectious individuals. Individuals are identified as cases if they experience one or more of the following symptoms: earache, runny nose, sore throat, hoarseness, cough, phlegm, wheezing, or painful breath. Attack rates tend to decrease with increasing age. Several factors may explain this trend. First, susceptibility probably decreases with increasing age, i.e., antibody acquisition. Second, there is closer physical contact and less hygienic behavior among young children who congregate in mixing groups such as preschools and daycare centers. Finally, the virus exhibits decreasing pathogenicity with increasing age (Gwaltney, 1982). Previous studies of rhinovirus infection reported by Cate, Couch, and Johnson (1964) and Douglas (1970) have suggested that reasonable values for the mean lengths of the latent (T_1), incubation ($T_1 + T_2$), and infectious periods ($T_2 + T_3$) are approximately 2, 2.5, and 12 days, respectively. The distributions of T_1, T_2, and T_3, which are given in Table 1, are calculated from these mean values. The Tecumseh study was designed to evaluate the effects of intranasal interferon spray in preventing rhinovirus infection. Households were randomized to receive either interferon or a placebo, using a method of post-exposure prophylaxis that had previously achieved significant reduction in transmission (Douglas et al., 1985; Hayden et al., 1986). However, the dose was approximately half of that given to individuals in earlier studies. At this level, no significant differences were found (Monto et al., unpublished manuscript). Therefore, the interferon prophylaxis and placebo households have been combined.

Table 1
Probability distributions[a] *used for the random variables T_1, T_2, and T_3 in the analysis of the rhinovirus data*

t (days)	$\Pr(T_1 = t)$	$\Pr(T_2 = t)$	$\Pr(T_3 = t)$
0	.00	.50	.00
1	.40	.40	.00
2	.35	.10	.00
3	.15		.00
4	.10		.00
5			.01
6			.03
7			.03
8			.03
9			.05
10			.10
11			.20
12			.20
13			.20
14			.10
15			.05
Mean	1.95	.60	11.48

[a] *Sources:* Cate et al. (1964) and Douglas (1970).

Table 2
Distribution of individuals by age group and household size and age-specific attack rates from the 1983 *study in Tecumseh, Michigan*

	Household size							
Age group	3	4	5	6	7	8	9	Attack rate
0–4	0	23	27	6	1	0	3	70%
5–17	4	57	82	25	13	4	9	57%
18+	2	76	71	17	7	4	6	51%

In the study, 199 households were monitored for viral respiratory infection. To be eligible for recruitment, each household had to have at least two members eligible for spraying and at least two children under age 12. Throughout the study, families recorded the presence or absence of common cold symptoms daily; thus, the time units used in the analysis are days. A specimen for viral culture was obtained within 24 hours after the onset of symptoms from all individuals who felt they had a cold. Cultures were also obtained from apparent secondary cases. Standard methods were used for virus isolation.

Any person reporting symptoms in an uninfected household is defined to be an index case. A household was considered to be uninfected if either no other household members had been identified, or more than $\max\{T_1 + T_2 + T_3\}$ days had passed since the last individual was identified. Then all household members were considered to be susceptible again. Because symptom data are rarely 100% sensitive and specific for infection, we chose to use only those illnesses occurring within 3 weeks of illness onset in an index case who had a positive culture for rhinovirus. Although some rhinovirus episodes are missed using this procedure, we are assured that persons displaying symptoms subsequent to a rhinovirus positive index infection were in fact exposed to rhinovirus. The deletion of uninvaded households and households having no rhinovirus positive index cases does not affect the estimates of the SARs. The final data set consists of 91 households containing 437 individuals. These households were observed for lengths of time which varied between 71 and 99 days.

As an illustration, age is considered to be a risk factor. First, we model age-specific variation in susceptibility using three age groups: 0–4 years (preschool), 5–17 years (school age), and 18+ years (adult). The distribution of individuals by age group and household size is given in Table 2. The age-specific attack rates are also given in Table 2 and are calculated as the number of individuals identified as cases in each age group divided by the number of individuals in that age group. As expected, the susceptibility decreases with increasing age as individuals develop immunity due to repeated rhinovirus infections. Since uninvaded households have been deleted from the analysis, we expect the CPIs to be overestimated. However, the relative differences between risk groups should remain fairly stable. The distribution of households by size and number of cases is given in Table 3. The CPI estimates in Table 4 are based on 56 days of exposure, the average length of a rhinovirus season in Tecumseh (Longini et al., 1984). These estimates decrease from .432 for preschool children to .243 for adults. These figures indicate that a preschool child has about a 43% chance of being infected after 56 days of exposure to community sources of infection, while an adult has only a 24% chance of being infected after the same exposure. A similar pattern is observed for the SARs, which decrease from 18.5% for preschool children to 10.8% for adults. Thus, a preschool child has an 18.5% chance of being infected by a single infectious household member, while an adult has only a 10.8% chance of being infected by a single infectious household member. The hypothesis tests in Table 4 indicate a significant difference among the three CPIs. Pairwise comparisons reveal that the 18+ age group differs

Table 3
Distribution of households by size and number of cases from the 1983 *study in Tecumseh, Michigan*

Total number of cases[a]	Household size						
	3	4	5	6	7	8	9
1	1	11	9	2	1	1	0
2	0	9	3	1	0	0	0
3	1	9	6	2	0	0	0
4	0	2	5	0	0	0	0
5	0	3	5	0	1	0	1
6	0	4	5	0	0	0	1
7 or more	0	1	3	3	1	0	0
Total	2	39	36	8	3	1	2

[a] Note that one individual could become a case more than once during the observation period.

Table 4
Estimated CPIs and SARs ± one standard deviation for the rhinovirus epidemic season (1983) *in Tecumseh, Michigan, with age as a risk factor for susceptibility*

Age	CPI[a]	SAR
0–4	.432 ± .060	18.5 ± 4.37
5–17	.353 ± .035	14.1 ± 2.45
18+	.243 ± .035	10.8 ± 2.01
$H_{0,\mathrm{CPI}}$: $\mathrm{CPI}_{0-4} = \mathrm{CPI}_{5-17} = \mathrm{CPI}_{18+}$		$H_{0,\mathrm{SAR}}$: $\mathrm{SAR}_{0-4} = \mathrm{SAR}_{5-17} = \mathrm{SAR}_{18+}$
χ^2 (2df) = 9.325 (P = .009)		χ^2 (2df) = 3.012 (P = .222)

[a] Based on 56 days of exposure.

Table 5
Estimated CPIs and SARs ± one standard deviation for the rhinovirus epidemic season (1983) *in Tecumseh, Michigan, with age as a risk factor for susceptibility and infectiousness*

Age			
Suscept.	Inf.	CPI[a]	SAR
0–17	any	.370 ± .031	—
18+	any	.243 ± .034	—
0–17	0–17	—	17.4 ± 3.22
0–17	18+	—	12.5 ± 3.76
18+	0–17	—	10.7 ± 2.42
18+	18+	—	10.9 ± 5.30
H_{0I}: $\mathrm{SAR}_{0-17,0-17} = \mathrm{SAR}_{0-17,18+}$ $\mathrm{SAR}_{18+,0-17} = \mathrm{SAR}_{18+,18+}$		H_{0S}: $\mathrm{SAR}_{0-17,0-17} = \mathrm{SAR}_{18+,0-17}$ $\mathrm{SAR}_{0-17,18+} = \mathrm{SAR}_{18+,18+}$	
χ^2 (2df) = .845 (P = .665)		χ^2 (2df) = 3.080 (P = .214)	

[a] Based on 56 days of exposure.

significantly from the 0–4 and the 5–17 age groups at the $\alpha = .05$ level. These results are as expected since it is presumed that adults may build up immunity to some of the rhinovirus strains, and hence are at less risk of infection when exposed. Although the null hypothesis $H_{0,\mathrm{SAR}}$ is not rejected ($P = .222$), it appears that there is a decline in SAR with increasing age.

Finally, we classify individuals into two age groups (0–17 and 18+) to illustrate the use of the model to explore the dependence of susceptibility and infectiousness on age. Younger individuals may be more infectious to others than older individuals because of their less hygienic personal habits. Table 5 presents the estimated CPIs by the age group of

the susceptible and the SARs by the age groups of the susceptible and the infectious individuals. As before, there is a decrease in the CPI from .370 for children to .243 for adults ($P = .006$). The highest SAR is estimated for a susceptible child exposed to an infectious child, SAR = 17.4%. Thus, an infected child will infect a susceptible child in the same household about 17% of the time. The lowest SAR is estimated to be 10.7% for a susceptible adult exposed to an infectious child. In addition, infected adults appear to be slightly more infectious to children (SAR = 12.5) than to other adults (SAR = 10.9). The differences between pairs of SARs are not statistically significant.

6. Discussion

The incidence data model presented in this paper has been shown to be useful in describing the spread of an infectious disease throughout a community and within a household. The probability that an individual is infected from the community or from another household member can be estimated jointly from the illness incidence data, and individuals may differ with regard to susceptibility or infectiousness. The incidence data model is sufficiently general to model epidemics for diseases that confer no immunity, temporary immunity, or permanent immunity following infection. Individuals who become infected may be returned to the set of susceptibles, and after a sufficient period of time they may be identified again. The smallest possible number of time units between two successive identifications of the same individual is specific for a given infectious agent and equals the length of time needed for an individual to pass through states 3, 4, 1, and 2, in that order (see Section 2). In practice, successive identifications not separated by $\max\{T_3 + t_4 + T_1 + T_2\}$ time units are assumed to belong to a single prolonged episode.

Although incidence data are more difficult to collect than final attack rate data, more information about the nature of the spread of a particular infectious agent can be obtained from incidence data. With incidence data, the order in which members of a given household are identified is known. Therefore, the likelihood function can be constructed using more information. Cofactors (risk levels) can be allowed to vary over time, changing between time units when necessary but not changing within any given time unit. Identification of clustering of cases is also possible with incidence data. Another advantage of this model is that households do not have to remain under observation for the same amount of time. Furthermore, if individuals within the household do not all live in the household for the same time period, the likelihood function can still be constructed as in (2.1)–(2.3). With this approach it is assumed that an individual who enters the study late enters as either a susceptible individual or an identified individual, provided the number of time units that have passed since the individual was identified is known. In addition, we assume that all new arrivals occur at the start of a time period. Similarly, a person may leave the household at any time, either temporarily or permanently.

Additionally, we assume that the per-time-unit probability that an individual at susceptibility level r escapes infection from the community, b_r, is constant over the length of the observation period. However, since b_r is a per-time-unit probability, it is within the scope of the model to allow b_r to vary over time, in which case it is written as $b_{r(t)}$. Of course, the functional form of $b_{r(t)}$ must be specified either from prior knowledge of the disease prevalence in the community or from a separate sample designed to estimate the functional form of $b_{r(t)}$. If b_r is written as some function of time plus one or more unknown parameters, then $b_{r(t)}$ can be estimated indirectly by first estimating the additional nuisance parameters. However, since our primary interest in this paper is to estimate q_{rh}, we have chosen to make the simplifying assumption that b_r is constant over the time interval $t = 0$ to $t = T$ for the purpose of presenting the general concepts of this model.

These advantages of incidence data analysis do not come without a price. For one to use the above model to analyze such data, one must be able to specify the probability

distributions of the random variables T_1, T_2, and T_3 defined above. A complete description of these distributions may not be available. Furthermore, one must usually rely on symptom data, which are rarely 100% sensitive and specific for infection, rather than on more precise ascertainment techniques such as serology, which cannot be obtained for each time unit. As an example, cost and/or participant well-being would preclude taking blood samples on a daily basis. A study design that combines the advantages of incidence and final attack rate data is a design that yields periodic incidence data. This design requires that individuals be sampled at periods shorter than those for final attack rate data, but longer than those for incidence data. Generally, accurate infection information can be gained along with partial information on the time-dependent sequence of infections. Longini et al. (1989) have developed a model for the analysis of periodic incidence data in the transmission of the AIDS virus.

The results from the rhinovirus example demonstrate the usefulness of the incidence model. Researchers have attempted to calculate SARs from illness onset dates and periods of illness duration (see Longini, 1986, p. 98). These researchers assume that generations of cases would appear in serial intervals separated by the approximate length of the incubation period, i.e., $T_1 + T_2$. In this way, it is expected that secondary cases can be separated from later generations of cases. In addition, subsequent (to the index case) introductions would perhaps be dispersed such that they could be clearly identified as not being part of any intrahousehold generation of cases. However, the fact that the infectious period for rhinovirus (average of 12 days) is longer than the incubation period (average of 2.5 days) makes it highly unlikely that generations of cases would be discernible [see Foy et al. (1988) for further discussion on attempts to identify such generational intervals for cases involving rhinoviruses]. Furthermore, it is impossible to determine whether an individual was infected from another household member or from outside of the household, even if cases were to appear in orderly serial intervals. The model given here provides a practical method for estimating SARs and CPIs from illness incidence data without trying to identify generations of cases. A satisfactory goodness-of-fit statistic for this model has not yet been developed, although Rampey (unpublished dissertation cited previously) presents possible directions for developing such a statistic.

With regard to rhinovirus transmission, it is interesting to note that although no statistically significant differences were found among the estimated SARs in Table 4, the decreasing SARs associated with increasing age confirm what has been reported in other studies (see Fox, Cooney, and Hall, 1975; Longini et al., 1984). Furthermore, the relative relationships of the point estimates of the SAR_{rh}s in Table 5 are similar to those that have been observed in studies of influenza A(H3N2) transmission (Addy, unpublished dissertation cited previously). Note from Table 5 that the highest secondary spread is among children exposed to infectious children. In addition, children seem to be at slightly greater risk of being infected by adults than adults are by children. The SARs for adults exposed to infectious children and to other adults are similar, again suggesting that adults may have developed some limited immunity to certain viruses. Another way of expressing the observed differences here is by calculating risk ratios. For example, when comparing children exposed to infectious children with adults exposed to infectious children, the risk ratio, $RR = (SAR_{0-17,0-17})/(SAR_{18^+,0-17})$, is estimated to be 1.62, again suggesting that children are at greater risk than adults of being infected by children.

The SAR estimates reported here are the first such risk-specific estimates obtained from infectious disease incidence data that take both infectiousness and susceptibility into account. Estimates such as these should be useful to researchers studying the spread of an infectious disease. Estimates of risk-specific SARs can be used to identify individuals likely to be infected if exposed. These individuals could then take appropriate steps to reduce the

risks of becoming infected. The model presented here can also be used to identify individuals who, when infected, are more infectious to others. This information is important for planning intervention strategies, such as vaccination of selected groups in a population.

ACKNOWLEDGEMENTS

This research was done in partial fulfillment of the requirements for the first author's Ph.D. in statistics and biometry at Emory University. The research was partially supported by NIH Grant 1-R01-AI22877.

RÉSUMÉ

Un modèle par intervalle est proposé pour analyser la distribution dans le temps des cas de maladies infectieuses dans un échantillon de familles. Lorsqu'un cas est identifié, c'est à dire quand les symptomes de la maladie apparaissent, les probabilités d'avoir été contaminé soit par l'un des membres de la famille, soit par la collectivité sont estimées en fonction de divers facteurs de risques. La méthode du maximum de vraisemblance est utilisée pour estimer les paramètres du modèle. Un sujet peut être classé en fonction de son niveau de susceptibilité individuelle et des caractéristiques épidémiologiques de l'infection. Le modèle est adjusté à partir des données cliniques et virologiques recueillies lors d'une épidémie due à un rhinovirus survenue à Tecumseh, Michigan. On observe en particulier une diminution du risque d'infection lorsque l'âge des sujets augmente.

REFERENCES

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

Cate, T. R., Couch, R. B., and Johnson, K. M. (1964). Studies with rhinovirus in volunteers: Production of illness, effect of naturally acquired antibody, and demonstration of a protective effect not associated with serum antibody. *Journal of Clinical Investigation* **43,** 56–67.

Douglas, R. G., Jr. (1970). Pathogenesis of rhinovirus common colds in human volunteers. *Annals of Otology, Rhinology, and Laryngology* **79,** 563–571.

Douglas, R. M., Albrecht, J. K., Miles, H. B., Moore, B.W., Read, R., and Workwick, D. A. (1985). Intranasal interferon-α_2 prophylaxis of natural respiratory infection. *Journal of Infectious Diseases* **151,** 731–736.

Fox, J. P., Cooney, M. K., and Hall, C. E. (1975). The Seattle virus watch. V. Epidemiological observations of rhinovirus infections, 1965–1969, in families with young children. *American Journal of Epidemiology* **101,** 122–143.

Foy, H. M., Cooney, M. K., Hall, C., Malmgren, J., and Fox, J. P. (1988). Case-to-case intervals of rhinovirus and influenza virus infections in households. *Journal of Infectious Diseases* **157,** 180–182.

Grizzle, J. E., Starmer, C.F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25,** 489–504.

Gwaltney, J. M. (1982). Rhinoviruses. In *Viral Infections of Humans*, 2nd edition, A. S. Evans (ed.). New York: Plenum Medical.

Haber, M., Longini, I. M., Jr., and Cotsonis, G. A. (1988). Statistical analysis of infectious disease data via log-linear models. *Biometrics* **44,** 163–173.

Hayden, F. G., Albrecht, J. K., Kaiser, D. L., and Gwaltney, J. M., Jr. (1986). Prevention of natural colds by contact prophylaxis with intranasal alpha$_2$-interferon. *New England Journal of Medicine* **314,** 71–75.

IMSL, Inc. (1987). *STAT/LIBRARY User's Manual, Version 1.0.* Houston, Texas: IMSL.

Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics, Volume 1,* 4th edition. New York: Macmillan.

Longini, I. M., Jr. (1986). Modeling influenza epidemics. In *Options for the Control of Influenza,* A. P. Kendal and P. A. Patriarca (eds), 89–105. New York: Alan R. Liss.

Longini, I. M., Jr., Clark, W. S., Haber, M., and Horsburgh, C. R., Jr. (1989). The stages of HIV infection: Waiting times and infection transmission probabilities. *Lecture Notes in Biomathematics* **83,** 112–137.

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

Longini, I. M., Jr., Koopman, J. S., Haber, M., and Cotsonis, G. A. (1988). Statistical inference for infectious diseases: Risk-specific household and community transmission parameters. *American Journal of Epidemiology* **128,** 845–859.

Longini, I. M., Jr., Koopman, J. S., Monto, A. S., and Fox, J. P. (1982). Estimating household and community transmission parameters for influenza. *American Journal of Epidemiology* **115,** 736–751.

Longini, I. M., Jr., Monto, A. S., and Koopman, J. S. (1984). Statistical procedures for estimating the community probability of illness in family studies: Rhinovirus and influenza. *International Journal of Epidemiology* **13,** 99–106.

Received November 1988; *revised July and November* 1990; *accepted December* 1990.