AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 128, No. 4
Printed in U.S.A.

STATISTICAL INFERENCE FOR INFECTIOUS DISEASES

RISK-SPECIFIC HOUSEHOLD AND COMMUNITY TRANSMISSION PARAMETERS

IRA M. LONGINI, JR.,[1] JAMES S. KOOPMAN,[2] MICHAEL HABER,[1] AND GEORGE A. COTSONIS[1]

Longini, I. M., Jr. (Dept. of Epidemiology and Biostatistics, Emory U., Atlanta, GA 30322), J. S. Koopman, M. Haber, and G. A. Cotsonis. Statistical inference for infectious diseases: risk-specific household and community transmission parameters. *Am J Epidemiol* 1988;128:845–59.

A statistical model is presented for the analysis of infectious disease data from family studies in the community. The model partitions the sources of infection into those from within the household and those from the community at large. The parameters reflecting these sources of infection are estimated as functions of the risk factors. This new model is used to overcome problems associated with the lack of independence of observations in infectious disease data and negative confounding due to the association of unmeasured exposures and immunity. An example of how this new statistical model is used to provide a clearer and less confounded description of risk factor effects is presented for data from influenza A(H3N2) epidemic seasons in the Tecumseh Respiratory Illness Study. The risk factors examined are age and pre-epidemic season antibody level as measured by the hemagglutination-inhibition test, while the outcome is the infection rate. A standard analysis of the data indicates that the efficacy of protective antibodies is 70% in children and only 47% in adults. However, such an efficacy measurement is negatively confounded by past exposure which is age dependent. By means of the model, the true, unconfounded, efficacy of protective antibodies is shown to be 90% in both adults and children.

antibodies; biometry; disease outbreaks; orthomyxoviridae; vaccines

A major goal of infectious disease epidemiology is to assess the impact of important risk factors on the probability of infection for persons. Such risk factors generally affect the probability that an infectious agent will be successfully transmitted to susceptible persons. However, the usual method of regressing infection or illness attack rates against the levels of various measured risk factors, e.g., contingency tables, log-linear models, and logistic regression, can only provide indirect and sometimes ambiguous information about the relation between transmission probabilities and risk factors. This problem arises because each time a person is infected, the risk of infection for those persons close to him or her increases directly due to potential transmission of the infectious agent. In addition, the use of standard statistical methods leads to markedly biased point estimates of absolute or relative risks because current

Received for publication June 26, 1987, and in final form November 24, 1987.

[1] Division of Biostatistics, Department of Epidemiology and Biostatistics, Emory University, Atlanta, GA 30322. (Reprint requests to Dr. Ira M. Longini, Jr. at this address.)

[2] Department of Epidemiology, University of Michigan, Ann Arbor, MI.

This research was supported by NIH grant 1-R01-AI22877 and General Clinical Research Center Public Health Research Grant 5-M01-RR000039.

The authors thank Dr. Kevin Sullivan for his comments and Dr. Arnold Monto for making the Tecumseh data available.

risk factors for exposure to infection are often correlated with past such exposures and, thus, with acquired immunity. Since the risk factors associated with past exposure can never be fully specified and entered into the analysis, the confounding effect of these past exposures on the relation between present risk factors and disease outcome can not be controlled for via stratification, log-linear models, or logistic regression.

In this paper, a statistical procedure is used to help resolve some of these difficulties by relating risk factors directly to the two major sources of infection: 1) those within the household and 2) those from the community at large. The model used in this paper is an extension of the probabilistic model developed by Longini and Koopman (1) and Longini et al. (2). In that model, household and community sources of infection are represented by the two parameters defined as the household secondary attack rate and the community probability of infection, respectively. This model has been used to measure and compare the transmissibility of different types and subtypes of influenza in the household and the community, both in Tecumseh, Michigan and Seattle, Washington (2, 3). In addition, the model has been applied to the analysis of data from rhinovirus (4), rotavirus (5), and dengue fever (6) epidemics. The modeled estimator for secondary attack rate has been shown to be robust under a number of common simulated field conditions (7). The extension presented here of that model consists of specifying the secondary attack rate and community probability of infection in the basic model of Longini et al. (1, 2) at different levels of risk factors. Technical details can be found in Haber et al. (8). In this paper, the basic technique is introduced as a new tool for assessing the impact of risk factors on infection transmission.

The model

Probability model

The model centers on the probability that a particular pattern of infection occurs in a household during the course of an epidemic, given an initial pattern of risk for the persons in the household. The model is formulated for a single risk factor which can be categorized into $r = 1, 2, \ldots, R$ levels of risk. This formulation can be generalized to multiple risk factors whose joint exposure define specific risk categories. Define Q_r as the probability that a susceptible person, at the rth risk factor level, escapes being infected by a *single* infected household member during the latter's entire infectious period. Further, let B_r be the probability that the susceptible person, at the rth risk factor level, escapes being infected from community sources during the course of the epidemic period. Now consider a household with s initial susceptibles whose risk levels are $r_1, r_2, \ldots, r_s$. Then index the s susceptibles in a household by $i = 1, 2, \ldots, s$. In order to describe the pattern of infection in a household, define the indicator variable x_i, where $x_i = 1$ if the ith household member is infected and $x_i = 0$ if not infected. Then $k = \sum_{i=1}^{s} x_i$ is the total number of persons infected in a household. In order to derive the probability of a particular pattern of infection $(x_1, \ldots, x_s)$ given a particular pattern of risk $(r_1, \ldots, r_s)$, for a household with s initial susceptibles, the following assumptions are made:

1) The infection escape parameters Q_r and B_r are independent of the household size unless household size is specified as a risk factor.

2) Every infected person is equally infectious to other persons regardless of that infected person's risk factor level. (Of course, the susceptibility and degree of contact with infectives varies for susceptibles according to their risk factor level, r.)

3) Household members mix at random within the household.

4) The event that a person is infected from the community is independent of the number of susceptible and infected persons in his or her household.

5) Following infection, a person becomes immune for the remainder of the epidemic period.

Given the above five assumptions, the probability of interest is given by

$$P(x_1, \ldots, x_s \mid r_1, \ldots, r_s) = \begin{cases} \prod_{i=1}^{s} B_{r_i}, & \text{for } k = 0, \\ P(\mathbf{1}_k \mid r_{j_1}, \ldots, r_{j_k}) \prod_{i:x_i=0} B_{r_i} Q_{r_i}^{k}, & \text{for } k = 1, \ldots, s-1, \\ 1 - \sum\limits_{x_1, \ldots, x_s : \sum x_i < k} P(x_1, \ldots, x_s \mid r_1, \ldots, r_s), & \text{for } k = s, \end{cases} \tag{1}$$

where $j_1, \ldots, j_k$ are the indices of the k infected persons and $\mathbf{1}_k$ denotes the array $(1, \ldots, 1)$ of order k. The derivation of equation 1 is given in the Appendix. There are R^s possible risk factor patterns and 2^s possible infection patterns for each risk factor pattern, yielding $2^s \times R^s$ probabilities from equation 1. For example, suppose $R = 2$ and $s = 2$, then there are $2^2 \times 2^2 = 16$ probabilities. Eight of the possible outcomes are listed below:

$$\begin{aligned} P(0,0 \mid 1,2) &= B_1B_2 = P(0,0 \mid 2,1), \\ P(1,0 \mid 1,2) &= P(1 \mid 1)B_2Q_2 = (1-B_1)B_2Q_2 = P(0,1 \mid 2,1), \\ P(0,1 \mid 1,2) &= P(1 \mid 2)B_1Q_1 = (1-B_2)B_1Q_1 = P(1,0 \mid 2,1) \\ P(1,1 \mid 1,2) &= 1 - P(0,0 \mid 1,2) - P(1,0 \mid 1,2) - P(0,1 \mid 1,2) \\ &= 1 - B_1B_2 - (1-B_1)B_2Q_2 - (1-B_2)B_1Q_1 = P(1,1 \mid 2,1). \end{aligned}$$

The model given by equation 1 reduces to the model of Longini et al. (1, 2) when the risk factors are on a household level. In general, the $2R$ model parameters ($B_1, B_2, \ldots, B_R, Q_1, Q_2, \ldots, Q_R$) are estimated by fitting the probability model given by equation 1 to household level infection data as described below.

Data and parameter estimation

The data are taken from family studies in the community where the antibody level of all persons in households can be ascertained both before and after the epidemic period. The initial antibody level is used to ascertain the susceptibility of each person and the paired sera are used to identify those persons who were infected during the course of the epidemic. In addition, the levels of the risk factors of interest are determined for each person. The likelihood function for model 1 is derived as the product of the probabilities of the observed outcomes for all the households in the sample. Numerical methods are used to calculate the maximum likelihood estimates for the parameters. Hypothesis testing is carried out in a standard fashion (see Haber et al. (8) for details).

In the case when the risk factors are measured on the household level, the probability model given in equation 1 can be transformed into a log-linear model (8). Then, the parameters B_r and Q_r can be estimated using the weighted least squares method for log-linear models (9).

Risk-specific secondary attack rate and community probability of infection

The parameter $1 - Q_r$ is the probability that a single infected household member will infect another susceptible household member who is at the rth level of the risk factor. Thus, following Longini et al. (2) and Haber et al. (8), the risk-specific secondary attack rate is given by $\text{SAR}_r = (1 - Q_r) \times 100$. The parameter $1 - B_r$ is the probability that a susceptible person, at the rth level of the risk factor, will be infected from a community source during the course of the epidemic period. The risk-specific community probability of infection will be given by $\mathrm{CPI}_r = 1 - B_r$. The maximum likelihood estimates $\hat{Q}_r$ and $\hat{B}_r$ are used to provide estimates of SAR_r and CPI_r along with their standard errors. The hypothesis tests on the parameters B_r and Q_r can be used to test hypotheses concerning the risk ratios of the CPI_r's and SAR_r's. For example, the null hypothesis H_0: $Q_1 - Q_2 = 0$ is the same as the null hypothesis H_0: $\mathrm{SAR}_1 - \mathrm{SAR}_2 = 0$ or $\mathrm{SAR}_1/\mathrm{SAR}_2 = 1$. If several hypotheses are tested simultaneously, then the overall significance level can be controlled for using multiple-comparison techniques, e.g., the Bonferroni inequality.

Analysis of influenza data

To illustrate the methods described above, an analysis of household and community transmission dynamics of influenza A(H3N2) in Tecumseh, Michigan is given. The impact of two important risk factors (age and pre-season antibody level) on the transmission parameters is assessed.

The data

The data are taken from the Tecumseh Study of Respiratory Illness (10, 11). In this study, a 10 per cent cross-sectional (random) sample of households from Tecumseh's population were kept on report for the periods of 1965–1971 (11) and 1976–1981 (12). Data on personal and household level risk factors were collected for all participating household members. Blood specimens were collected every six months on a staggered basis, so that, in any month, approximately one-sixth of the population on report had blood obtained. The hemagglutination-inhibition antibody test was performed on the blood samples.

The influenza epidemic was defined each year using virus isolation and illness incidence information (10–12). Each epidemic period was bracketed by pre- and post-epidemic season bleedings. The pre-season antibody titer level was used to establish the susceptibility of each person. Susceptible persons were considered to have been infected during the course of the epidemic period if there was a "significant" rise in antibody titer when comparing pre- and post-season titers. In addition, a person was considered to have been infected if virus was isolated from that person.

Data were combined from two influenza A(H3N2) epidemics which occurred during the 1977–1978 and 1980–1981 epidemic seasons in Tecumseh. Table 1 shows the epidemic infection attack rates for persons with different pre-season titer levels (1 : x) to influenza A(H3N2). Persons with pre-season titer levels $x < 8$ had considerably higher attack rates than those with levels $8 \leq x \leq 64$, while those with levels $x \geq 128$ were not infected at all. Thus, pre-epidemic hemagglutination inhibition titer level was categorized into three risk groups according to pre-season antibody level: $x < 8$, low antibody level or no antibody (highly susceptible); $8 \leq x \leq 64$, higher antibody level (less susceptible); $x \geq 128$, high antibody level (immune). Table 2 shows the attack rates for persons in different age groups. Clearly, attack rates decrease with increasing age. In order to ensure sufficient num-

TABLE 1

Infection rates by pre-season antibody titer: influenza A(H3N2) epidemic seasons 1977–1978 and 1980–1981 combined, in Tecumseh, Michigan

Pre-season antibody titer (1:x)	No. observed	Fraction infected
<8†	836	0.234
8	267	0.139
16	153	0.059
32	137	0.088
64‡	87	0.046
≥128	26	0
Total	1,506	0.171

† Persons with pre-season antibody titer $x < 8$ were categorized as having a low level of antibody.

‡ Persons with pre-season antibody titer $8 \leq x \leq 64$ were categorized as having a higher level of antibody, and those persons with pre-season antibody titer $x \geq 128$ were considered to be immune.

TABLE 2

Infection rates by age groups: influenza A(H3N2) epidemic seasons 1977–1978 and 1980–1981 combined, in Tecumseh, Michigan

Age group (years)	No. observed	Fraction with pre-season titer <8	Fraction infected
Preschool (0–3)	147	0.837	0.270
School (4–17)	361	0.490	0.227
Adult (18+)	998	0.537	0.139
Total	1,506	0.555	0.171

bers of persons in each risk category, age was categorized into two risk groups: 0–17 years (children), 18+ years (adults). The fraction of each age group that had low pre-season antibody levels is also shown in table 2. Over 80 per cent of preschool children had little or no antibody to influenza A(H3N2) before the epidemics, while around 50 per cent of the older persons had little or no antibody. The overall trend is for persons to have increasing antibody levels with increasing age due to past exposure to various strains of influenza A(H3N2). However, this increase may not be completely smooth because some protection is lost with age and the hemagglutination-inhibition test used to classify levels of antibody is most specific to certain current strains of influenza A(H3N2).

A standard method of organizing these data is in the form of a 2 × 2 × 2 table (13). The data in table 3 are organized in such a fashion by three factors: infected vs. not infected, child vs. adult, and low vs. higher pre-season titer. From an examination of the attack rates, both young age and low pre-season antibody level increase the risk of infection. The relative effect of each particular factor is greater in the presence of the other factor. The joint effects are significantly greater than multiplicative ($p < 0.0001$). Both the point estimates of the risk ratios and the statistical tests used in table 3 may be misleading since persons are clustered into households and there is transmission of infection within these households. Thus, there is lack of independence in the response variable (infected or not infected) on the household level and the sampling structure of the table is not multinomial. The estimated variance for the estimator of the attack rate will be lower than the actual sample variance, resulting in increased probabilities of Type I errors during hypothesis testing (14). The probability model 1 is formulated on a household level. Thus, the household level sampling structure is incorporated directly into the analysis.

A major problem in interpreting the results from an analysis like that presented in table 3 is that the association of current exposure with past exposure will create a strong source of negative confounding. Persons with higher pre-season antibody levels have higher levels because they had a higher probability of encountering influenza infection during their normal activities than other persons. Those "usual activities" are unlikely to have changed and, thus, persons with higher pre-season antibody levels are likely to be more exposed to infection than persons with low pre-season antibody levels. If this source of negative confounding had been greater in adults than in children, as seems likely to be the case, such confounding could have created the greater than multiplicative relations seen in table 3.

The standard analysis on the attack rates presented above can not be used to distinguish between the effects of extrahousehold and intrahousehold sources of infection on the risk of infection, but the probability model given by equation 1 can be used to separate these two sources of risk. In addition, the secondary attack rate is less subject to the source of negative confounding described above.

The fitted probability model

The probability model was fitted to the pooled data from the two influenza A(H3N2) epidemic seasons. First, the model was used to examine the crude effect of the pre-season antibody level on the

TABLE 3

Infection attack rates by pre-season antibody titer level stratified by age group: influenza A(H3N2) epidemic seasons 1977–1978 and 1980–1981 combined, in Tecumseh, Michigan

Pre-season antibody titer (1:x)	Infection status			Attack rate†	Risk ratio‡
	No. infected	No. not infected	Total		
Children (0–17 years)					
Low level ($x < 8$)	100	200	300	0.333	3.330*
Higher level ($8 \leq x \leq 64$)	20	180	200	0.100	
Total	120	380	500§	0.240	
Adults (18+ years)					
Low level ($x < 8$)	96	440	536	0.179	1.884*
Higher level ($8 \leq x \leq 64$)	42	402	444	0.095	
Total	138	842	980§	0.141	

* $p < 0.001$.

† Attack rate = no. infected/no. at risk.

‡ Risk ratio = ratio of the attack rates.

§ The total of 1,480 persons does not include the 26 "immune" persons.

The risk ratios across levels of age or across levels of pre-season antibody titer are different ($p < 0.0001$) using the chi-square test for lack of interaction.

community probability of infection and secondary attack rate, when the data were collapsed across age. Then, the crude effect of age on the community probability of infection and secondary attack rate was examined, when the data were collapsed across pre-season antibody level. Finally, the joint effects of pre-season antibody level and age on the community probability of infection and secondary attack rate were examined.

The fit of the probability model to the pooled data stratified on pre-season antibody titer level is shown in table 4. For example, from table 4, there were 63 households with a single susceptible at the low pre-season titer level and no susceptibles at the higher level. That susceptible person was not infected in 45 of such households, and he or she was infected in the other 18 households. There were 70 households with a single susceptible at the higher pre-season titer level and no susceptible at the low level. In this case, the person was not infected in 65 of the households but he or she was infected in five of the households. The estimated parameters are given at the bottom of the table. The expected frequencies are found by evaluating equation 1 at the parameter estimates. There is no pattern of deviations in table 4 that would suggest that the model does not describe the underlying process and the chi-square goodness-of-fit statistic is nonsignificant ($p = 0.235$).

The CPI_1 is estimated to be 0.164 ± 0.015, indicating that a person with a low level of pre-season antibody had about a 16 per cent chance of being infected from the community during the epidemic period. In contrast, CPI_2 was estimated to be 0.092 ± 0.013, indicating that a person with a higher level of pre-season antibody had about a 9 per cent chance of being infected from the community. The risk ratio, when low to higher levels of pre-season antibody were compared, i.e., $RR = CPI_1/CPI_2$, was estimated to be 1.783 ($p < 0.0005$), indicating that a person with a low level of antibody was nearly twice as likely to be infected from the community as a person with a higher level.

The SAR_1 is estimated to be 26.0 ± 3.0, indicating that a person with a low level of pre-season antibody had about a 26 per cent chance of being infected by another household member during the course of the latter's infectious period, while SAR_2 was estimated to be 2.1 ± 2.6 for higher titers. The estimated risk ratio for secondary attack rates, when low to higher levels of pre-season antibody were compared, was found to be $\widehat{RR} = 12.4$ ($p < 0.0001$), indicating that higher levels of antibody are quite protective given exposure to a single infected person.

As pointed out above, age is a second important risk factor. The probability model was fitted to the pooled influenza A(H3N2) data stratified on age, with level 1 for children (ages 0–17 years) and level 2 for adults (age 18+ years). The estimated CPIs for children and adults were 0.175 ± 0.023 and 0.113 ± 0.012, respectively. When the risk of infection from the community for children to that for adults was compared, this yields an estimated risk ratio of 1.549 ($p < 0.01$). The estimated secondary attack rates were 22.2 per cent and 11.1 per cent for children and adults, respectively. Thus, the risk ratio was estimated to be 2.0 (not significant).

The association between age and pre-season antibody level, as well as the need to describe how the effects of these two factors interact, requires stratification of the low and higher pre-season antibody level by age group, as is done in table 3. This is carried out using the model 1 and by specifying the community probability of infection and secondary attack rate at four levels of risk defined by joint exposure to each of the dichotomous variables: age and pre-season antibody level. The parameters are expressed as SAR_{ij} and CPI_{ij}, where index i indicates age level and j indicates pre-season antibody level. When the model is fitted to the data, the fit is adequate as indicated by there being no discernible pattern of deviations, no dramatic differences in observed and expected frequencies in any category, and a nonsignificant chi-square goodness-of-fit statistic ($p = 0.361$). The estimated secondary attack rates and community probability of infections and their standard deviations for each of the four exposure categories are presented in table 5. For example, the estimated secondary attack rate for a child with low pre-season antibody level is $\widehat{\mathrm{SAR}}_{11} = 36.6 \pm 6.2$, while it is $\widehat{\mathrm{SAR}}_{12} = 3.4 \pm 4.7$ for a child with higher pre-season antibody level. Thus, the risk ratio, when low to high levels of pre-season antibody for children are compared, is estimated to be $\widehat{RR} = \widehat{\mathrm{SAR}}_{11}/\widehat{\mathrm{SAR}}_{12} = 10.8$ ($p < 0.01$).

An important observation from table 5 is that the relative effect of protective antibody is much greater when measured by the secondary attack rates than by the community probability of infections. In children, the effect of protective antibody on the community probability of infections is 2.5-fold, while it is 10.8-fold on the secondary attack rates. Similarly, in adults, the effect of protective antibody on the community probability of infections is 1.5-fold, while the effect of protective antibody on the secondary attack rates is 11.4-fold. Another important observation is that, while there is a difference between the relative effect of antibody on the community probability of infections of adults and children (1.472 vs. 2.457, $p = 0.054$), there is almost no difference between the relative effect of protective antibody on the secondary attack rates of children and that of adults (11.4 vs. 10.8). A final observation, from table 5, is that given the same category of protective antibody, children have a higher community probability of infection and secondary attack rate than adults. However, these differences are statistically significant ($p < 0.05$) only at the low level of pre-season antibody.

DISCUSSION

Infectious disease data, such as those presented above, have two important characteristics that should be taken into account in analysis. The first is that the data are frequently sampled in clusters, such as households, schools, or other groups. This characteristic is incorporated directly into the probability model 1 since the model is formulated for the pattern of infection in households. A second is that the responses, i.e., infected or not infected, are correlated within the clusters due to the infection process (i.e., the probability that a person

TABLE 4

Observed and expected frequencies for households by the number of susceptibles from the influenza A(H3N2) seasons 1977–1978 and 1980–1981 combined, in Tecumseh, Michigan, stratified by pre-season antibody titer level†*

No. susceptible/household		No. infected/household		No. of households	
Pre-season antibody titer		Pre-season antibody titer		Observed	Expected
Low	Higher	Low	Higher		
1	0	0	0	45	52.7
		1	0	18	10.3
0	1	0	0	65	63.6
		0	1	5	6.4
2	0	0	0	52	49.6
		1	0	11	14.4
		2	0	8	7.0
1	1	0	0	52	50.1
		0	1	2	3.8
		1	0	8	9.6
		1	1	4	2.5
0	2	0	0	45	42.9
		0	1	6	8.5(a)
		0	2	1	0.6(a)
3	0	0	0	17	16.9
		1	0	4	5.5
		2	0	3	3.9
		3	0	5	2.7
2	1	0	0	28	24.7
		0	1	1	1.3(b)
		1	0	6	7.0
		1	1	0	1.4(b)
		2	0	2	3.3
		2	1	2	1.2(b)
1	2	0	0	16	17.2
		0	1	6	2.5(c)
		0	2	0	0.1(c)
		1	0	2	3.2
		1	1	1	1.7(c)
		1	2	0	0.2(c)
0	3	0	0	11	11.2
		0	1	4	3.3(d)
		0	2	0	0.5(d)
		0	3	0	0.0(d)
4	0	0	0	16	13.7
		1	0	4	4.3
		2	0	6	3.5
		3	0	0	3.5
		4	0	2	3.0

3	1	0	0	13	13.8
		0	1	0	0.6(e)
		1	0	6	4.3
		1	1	1	0.6(e)
		2	0	1	3.0
		2	1	0	0.8(e)
		3	0	5	2.0(e)
		3	1	0	0.8(e)
2	2	0	0	11	11.5
		0	1	0	1.3(f)
		0	2	1	0.1(f)
		1	0	1	3.2
		1	1	3	1.2(f)
		1	2	1	0.1(f)
		2	0	3	1.5(f)
		2	1	0	1.0(f)
		2	2	0	0.2(f)
1	3	0	0	10	12.5
		0	1	5	2.7
		0	2	0	0.3(g)
		0	3	0	0.0(g)
		1	0	2	2.3(g)
		1	1	1	1.7(g)
		1	2	2	0.4(g)
		1	3	0	0.0(g)
0	4	0	0	10	8.2
		0	1	2	3.1(h)
		0	2	0	0.7(h)
		0	3	0	0.1(h)
		0	4	0	0.0(h)
5	0	0	0	3	2.4
		all other		3	3.6
4	1	0	0	2	2.7
		all other		4	3.3
3‡	2‡			4	
2	3	0	0	4	5.2
		all other		6	4.8
1‡	4‡			2	
0‡	5‡			3	
Total				567	

[*] Only households with five or fewer susceptibles are included. Outcomes (a)-(h): those outcomes with the same letter were pooled for the goodness-of-fit test.

† Point estimates and standard errors:

Low pre-season antibody titer $\widehat{\mathrm{CPI}}_1 = 0.164 \pm 0.015$ $\widehat{\mathrm{SAR}}_1 = 26.0 \pm 3.0$

Higher pre-season antibody titer $\widehat{\mathrm{CPI}}_2 = 0.092 \pm 0.013$ $\widehat{\mathrm{SAR}}_2 = 2.1 \pm 2.6$

Overall goodness-of-fit χ^2 (28 df)‡ $= 33.01$, $p = 0.235$

Tests: $H_0 : \mathrm{CPI}_1/\mathrm{CPI}_2 = 1$ $\widehat{\mathrm{RR}} = 1.783$, $p < 0.0005$

$H_0 : \mathrm{SAR}_1/\mathrm{SAR}_2 = 1$ $\widehat{\mathrm{RR}} = 12.4$, $p < 0.0001$

Abbreviations: CPI, community probability of infection; df, degrees of freedom; SAR, secondary attack rate; RR, risk ratio.

‡ These combinations were not included in the goodness-of-fit test.

TABLE 5

Comparison of CPI$_{ij}$ *and* SAR$_{ij}$ *from the influenza A(H3N2) epidemic seasons 1977–1978 and 1980–1981 combined, in Tecumseh, Michigan, stratified by age group and pre-season antibody titer*†

Age		Pre-season antibody titer (1:x)	
		($j = 1$) Low level ($x < 8$)	($j = 2$) Higher level ($8 \le x \le 64$)
($i = 1$) Children	$\widehat{\text{CPI}}$	0.231 ± 0.032	0.094 ± 0.028
(0–17 years)	$\widehat{\text{SAR}}$	36.6 ± 6.2	3.4 ± 4.7
($i = 2$) Adults	$\widehat{\text{CPI}}$	0.131 ± 0.018	0.089 ± 0.015
(18+ years)	$\widehat{\text{SAR}}$	18.2 ± 4.4	1.6 ± 3.7
Risk ratios for CPIs:			
$H_0 : \text{CPI}_{11}/\text{CPI}_{12} = 1$		$\widehat{\text{RR}} = 2.457^{**}$	
$H_0 : \text{CPI}_{21}/\text{CPI}_{22} = 1$		$\widehat{\text{RR}} = 1.472$	
$H_0 : \text{CPI}_{11}/\text{CPI}_{21} = 1$		$\widehat{\text{RR}} = 1.763^{*}$	
$H_0 : \text{CPI}_{12}/\text{CPI}_{22} = 1$		$\widehat{\text{RR}} = 1.056$	
	Interaction :	$H_0 : \text{CPI}_{11}\text{CPI}_{22}/\text{CPI}_{12}\text{CPI}_{21} = 1$	$\widehat{\text{RR}}/\widehat{\text{RR}}' = 1.670$ ($p = 0.054$)
Risk ratios for SARs:			
$H_0 : \text{SAR}_{11}/\text{SAR}_{12} = 1$		$\widehat{\text{RR}} = 10.8^{**}$	
$H_0 : \text{SAR}_{21}/\text{SAR}_{22} = 1$		$\widehat{\text{RR}} = 11.4^{**}$	
$H_0 : \text{SAR}_{11}/\text{SAR}_{21} = 1$		$\widehat{\text{RR}} = 2.0^{*}$	
$H_0 : \text{SAR}_{12}/\text{SAR}_{22} = 1$		$\widehat{\text{RR}} = 2.1$	
	Interaction :	$H_0 : \text{SAR}_{11}\text{SAR}_{22}/\text{SAR}_{12}\text{SAR}_{21} = 1$	$\widehat{\text{RR}}/\widehat{\text{RR}}' = 0.95$

* $p < 0.05$.

** $p < 0.01$.

† Abbreviations: CPI, community probability of infection; SAR, secondary attack rate; RR, risk ratio. Overall goodness-of-fit χ^2 (23 degrees of freedom) = 24.794, $p = 0.361$.

escapes being infected in a cluster depends on the infection status of the other persons in the cluster). The probability model also incorporates this characteristic directly into the analysis via the term for the escape probability $B_{r_i}Q^k_{r_i}$. Both of these characteristics violate the basic statistical assumptions lying behind the standard contingency table approach, such as that used in table 3, or behind other techniques, such as logistic regression.

An alternative parameterization of the parameters could be $B = \exp(\beta \boldsymbol{U})$ and $Q = \exp(\gamma \boldsymbol{U})$, where $\boldsymbol{U}$ is column vector of individual level risk identifiers and/or measured characteristics and β and γ are row vectors of coefficients for the effects of community and household exposure to infection, respectively. Such a parameterization could be more efficient in terms of estimation, but it does impose an extra assumption on how the measured risk factors relate to the escape probabilities. The above formulation could also be quite useful if $\boldsymbol{U} = \boldsymbol{u}$ were a continuous scalar variable, or a mixture of categorical and continuous variables.

There are additional and perhaps more compelling reasons beyond those cited above to use equation 1 for the analysis of infectious disease data rather than the standard analyses presented in table 3. Equation 1 is used to separate the effects of risk factors into those effects due to transmission of infection from the community and those due to transmission from a single infected household member. Such a separation gives the model two important advantages: it allows for better characterization of the effects of risk factors on the transmission of infectious agents, and it allows for control of a common source of confounding in infectious disease data.

With respect to the first advantage, if different agents were ranked by infectiousness in the household and the community, the rankings would probably be quite different. Influenza and rhinoviruses, for example, sweep rapidly through communities, but have surprisingly low household secondary attack rates (2, 12, 15, 16). Shigella, on the other hand, does not sweep through communities in as complete and rapid a fashion as influenza and rhinoviruses (17), but the secondary attack rate is commonly high for shigella (18). By describing the relations of the effects of risk factors on the community probability of infections and secondary attack rates, the transmission model could be used to characterize modes of infectious agent transmission and to devise effective intervention strategies.

With respect to the second advantage of the transmission model, an important source of confounding in epidemiologic studies of infectious diseases is the association between current exposure levels and past exposure levels. The normal activities of persons are often determined by current stable factors. Since past exposure levels are strongly related to current immunity and antibody levels, this association of current and past exposures can markedly diminish the perceived effect of risk factors when using a standard analysis to measure such effect. The relations of the community probability of infections estimated using equation 1 are as subject to this kind of bias as are the attack rates when using the standard approach. However, the secondary attack rates are used to standardize exposure to a single person in the household. Since such exposure is constant across different risk groups, association of past and current exposures cannot confound the relations between risk factors and the secondary attack rate. Consequently, the secondary attack rates show a markedly greater relative effect of protection from higher levels of pre-season antibody. The relative effects of pre-season antibody on the secondary attack rates are nearly identical for children and adults. This is logical given that the secondary attack rates standardize to unity the probability of exposure to infected household members for adults and children. This equality between relative effects of pre-season antibody on children and those on adults when the evaluation of the effects using the secondary attack rates supports the notion that the inequality of relative effects, when the community probability of infections or the attack rates are used, is due to differential biases between the two age groups.

The secondary attack rates are higher for children than for adults within both categories of pre-season antibody level. This cannot be explained by differences between exposure frequencies in children and that in adults because the secondary attack rate standardizes the exposure frequency. One possible explanation is that there is residual effect within the pre-season antibody level categories, and children have lower levels than adults within categories. However, no evidence was found to support this explanation when the data were examined more closely. The most likely explanation is that adults derive some added protection from unmeasured non-hemagglutination immunity such as cellular immunity.

Antibody efficacy can be assessed by using a formula similar to that used for vaccine efficacy for any of the three following indices: the attack rate, community probability of infection, or secondary attack rate. Then

$$\text{antibody efficacy (index)} = \left(1 - \frac{\text{index}_2}{\text{index}_1}\right) \times 100 = \left(1 - \frac{1}{\text{RR}_{12}}\right) \times 100,$$

where 1 = low pre-season antibody titer level ($x < 8$) and 2 = higher pre-season antibody level ($8 \leq x \leq 64$). The crude antibody efficacy and the antibody efficacies for children and adults are given in table 6. Note that the antibody efficacy is underestimated using the attack rate or the community probability of infection as the index. The confounding effect of exposure

TABLE 6

Comparison of estimated antibody efficacy using different indices computed from the influenza A(H3N2) epidemic seasons 1977–1978 and 1980–1981 combined, in Tecumseh, Michigan†*

Index	Estimated antibody efficacy: stratification by age		Crude antibody efficacy
	Children (0–17 years)	Adults (18+ years)	
AR	70.0	46.9	58.8
CPI	59.3	21.1	56.1
SAR	90.7	91.2	91.9

* Antibody efficacy (Index) = $(1 - (\text{Index}_2/\text{Index}_1)) \times 100$, where 1 = low pre-season antibody titer ($x < 8$) and 2 = higher pre-season antibody titer ($8 \leq x \leq 64$).

† Abbreviations: AR, attack rate; CPI, community probability of infection; SAR, secondary attack rate.

intensity is readily apparent, as it was when considering risk ratios. The underestimate is most severe for adults. When the secondary attack rate is used as the index, the antibody efficacy is estimated to be 91 per cent in both children and adults.

The probability model can be used to evaluate the effect of intervention measures on transmission parameters. For example, in vaccine field trials, the risk factor assignment would be $r = 1$ for unvaccinated persons and $r = 2$ for vaccinated persons. Vaccine efficacy could be measured using the attack rate, community probability of infection, or secondary attack rate, as an index, where,

$$\text{vaccine efficacy (index)} = \left(1 - \frac{\text{index}_2}{\text{index}_1}\right) \times 100 = \left(1 - \frac{1}{\text{RR}_{12}}\right) \times 100.$$

The usual measure of vaccine efficacy is provided by the attack rate, i.e., index = attack rate. However, this index can be positively biased as a measure of vaccine efficacy in communitywide vaccine trials since the attack rate in unvaccinated persons may be artificially reduced due to herd immunity effects resulting from vaccination. The above formula actually measures vaccine effectiveness when the attack rate is used since it constitutes a mixture of vaccine efficacy, i.e., direct protection of the vaccinated, and herd immunity in communitywide vaccine trials. The index based on the community probability of infection would be better than the one based on the attack rate for measuring vaccine effectiveness because the former controls for the confounding effect of household transmission of infection. When the secondary attack rate is used as the index, the above formula provides a true measure of vaccine efficacy because the secondary attack rate measures the probability of infection given exposure to a single infected person.

The probability model could be used to measure the efficacy and effectiveness of other intervention measures such as use of interferon, virucidal nasal tissues, or general improvements in the host environment. The index based on the secondary attack rates has been recently used by Longini and Monto (19) to evaluate the efficacy of virucidal nasal tissues in interrupting the transmission of influenzavirus in the household. The efficacy of the tissue use was estimated to be 38 per cent when using the index based on the secondary attack rate and controlling for tissue usage level (19).

Analysis using the probability model presented here could facilitate the use of observational studies for decisions on the selection of antigen mixing strategies in the design of molecularly engineered vaccines. Influenzaviruses and rotaviruses have multiple antigens which could be included in vaccines. Those antigens to which antibodies have a greater than additive effect in protecting the host could be profitably combined in a vaccine. From observational studies such as those discussed here, antibodies to different antigens could be analyzed just as age and pre-season antibody level were analyzed in table 5. The relations between the secondary attack rates and the levels of various types of antibody could be used to determine what mix of antigens should go into vaccines, as outlined by Koopman (20).

Although the probability model is used to extract information from the data in an efficient manner, difficulties may occur due to lack of sufficient data leading to sparse cells. In the analysis of the influenza A(H3N2) data above, there were few children with high levels of pre-season antibody. Because of this difficulty, it was necessary to combine the data from the two influenza A(H3N2) seasons 1977–1978 and 1980–1981 in order to stratify on both age and pre-season antibody levels. This was done after verifying that the community probability of infections and secondary attack rates from the two seasons were not significantly different. Because of the above described sparseness, it was necessary to pool outcomes in order to compute an overall goodness-of-fit chi-square statistic for the fit of the model to data. Examples of such pooling can be found in table 4, where cells with small expected frequencies were added together.

A number of extensions to the probability model are planned. The first involves the analysis of sequential data over time. Such data usually consist of illness onset dates and duration in small groups such as families or large groups such as schools and communities. Although a number of models have been developed to estimate a single infectious contact parameter within the group (21–25), there has been little progress in the development of methods for relating risk factors directly to the infectious contact parameter when analyzing sequential (usually symptom) data. On a small group level, such a model could be used to estimate Q_r and B_r based on the sequential data. However, such a model requires additional parameters describing the distribution of the length of the latent and infectious periods (25). Such information is not required for the model presented in this paper. The basic model for relating risk factors to the infectious contact rate is being developed for larger population groups ($n > 30$).

The probability model presented here provides a useful tool for relating risk factors to sources of infection in family studies using serologic data. Valuable epidemiologic insights can be gained from such studies even though the exact timing of infection is not available. It is hoped that the usefulness of such large scale serologic studies such as the Tecumseh Study of Respiratory Illness will be recognized, and that these studies will be funded despite their cost.

References

1. Longini IM Jr, Koopman JS. Household and community transmission parameters from final distributions in households. Biometrics 1982;38:115–26.
2. Longini IM Jr, Koopman JS, Monto AS, et al. Estimating household and community transmission parameters for influenza. Am J Epidemiol 1982;115:736–51.
3. Longini IM Jr, Koopman JS, Monto AS. Estimation procedures for transmission parameters from influenza epidemics: use of serological data. Vopr Virusol 1983;2:176–81. (In Russian.)
4. Longini IM Jr, Monto AS, Koopman JS. Statistical procedures for estimating the community probability of illness in family studies: rhinovirus and influenza. Int J Epidemiol 1984;13:99–106.
5. Koopman JS, Monto AS, Longini IM Jr. The Tecumseh study. XVI. Family and community sources of rotavirus infection. Unpublished manuscript.
6. Gomez HD, Koopman JS, Addy CL, et al. Dengue epidemics on the Pacific coast of Mexico. Int J Epidemiol 1988;17:178–86.
7. Longini IM Jr, Seaholm SK, Ackerman E, et al. Simulation studies of influenza epidemics: assessment of parameter estimation and sensitivity. Int J Epidemiol 1984;13:496–501.
8. Haber M, Longini IM Jr, Cotsonis GA. Models for the statistical analysis of infectious disease data. Biometrics 1988;44:163–73.
9. Grizzle JE, Starmer CF, Koch GG. Analysis of categorical data by linear models. Biometrics 1969;25:489–504.
10. Monto AS, Napier JA, Metzner HL. The Tecumseh study of respiratory illness. I. Plan of study and observations on syndromes of acute respiratory disease. Am J Epidemiol 1971;94:269–79.
11. Monto AS, Kioumehr F. The Tecumseh study of respiratory illness. IX. Occurrence of influenza in the community, 1966–1971. Am J Epidemiol 1975;102:553–63.
12. Monto AS, Koopman JS, Longini IM Jr. Tecumseh study of illness. XIII. Influenza infection and

disease, 1976–1981. Am J Epidemiol 1985;121:811–22.

13. Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis. Cambridge, MA: MIT Press, 1975.
14. Cochran WG. Sampling techniques. 3rd ed. New York: John Wiley and Sons, 1977.
15. Fox JP, Cooney MK, Hall CE, et al. Rhinoviruses in Seattle families, 1975–1979. Am J Epidemiol 1985;122:830–46.
16. Fox JP, Cooney MK, Hall CE, et al. Influenzavirus infections in Seattle families, 1975–1979. II. Pattern of infection in invaded households and relation of age and prior antibody to occurrence of infection and related illness. Am J Epidemiol 1982;116:228–42.
17. Blaser MJ, Pollard RA, Feldman RA. Shigella infections in the United States, 1974–1980. J Infect Dis 1983;147:771–5.
18. Wilson R, Feldman RA, Davis J, et al. Family illness associated with shigella infection. J Infect Dis 1981;143:130–2.
19. Longini IM Jr, Monto AS. Efficacy of virucidal nasal tissues in interrupting familial transmission of respiratory agents: a field trial in Tecumseh, Michigan. Am J Epidemiol 1988;128:639–44.
20. Koopman JS. Analyzing the joint effects of two antibodies and the design of molecularly engineered vaccines. J Theor Biol 1985;116:569–85.
21. Bailey NTJ. The mathematical theory of infectious disease and its applications. 2nd ed. New York: Hafner, 1975.
22. Becker N. Estimation for an epidemic model. Biometrics 1976;32:769–77.
23. Dietz K, Schenzle D. Mathematical models for infectious disease statistics. In: Atkinson AC, Feinberg SE, eds. A celebration of statistics: the ISI centenary volume. New York: Springer, 1985:167–204.
24. Saunders IW. An approximate maximum likelihood estimator for chain binomial models. Aust J Stat 1980;22:307–16.
25. Longini IM Jr. The generalized discrete-time epidemic model with immunity: a synthesis. Math Biosci 1986;82:19–41.
26. Ludwig D. Final size distributions for epidemics. Math Biosci 1975;23:33–46.
27. Ball F. A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. Adv Appl Prob 1986;18:289–310.

Appendix

Derivation of the probability model

Equation 1 in the text is derived here based on the five assumptions given in the text. For a household with s susceptibles, the conditional probability of an infection array $(x_1, \ldots, x_s)$, with $k = \sum x_i$ infectives, given a risk factor array $(r_1, \ldots, r_s)$ is calculated by partitioning the sample space into three sets of outcomes corresponding to the cases where $k = 0$, $1 < k < s$ and $k = s$, as described below:

Set 1: $x_1 = \cdots = x_s = 0$, i.e., no persons are infected, $(k = 0)$. The probability of this event is simply the probability that all s susceptible persons escape being infected from community sources, or simply

$$P(0, \ldots, 0 \mid r_1, \ldots, r_s) = \prod_{i=1}^{s} B_{r_i}. \tag{A1}$$

Set 2: $x_i = 1$ for at least one $i = 1, \ldots, s$ but not all $x_i = 1$ $(0 < k < s)$. Let $\mathscr{S}_0$ and $\mathscr{S}_1$ denote groups of persons for whom $x_i = 0$ and $x_i = 1$, respectively. The event that $x_i = 0$ for all $i \in \mathscr{S}_0$ is denoted by E, and the event that $x_i = 1$ for all $i \in \mathscr{S}_1$ is denoted by F. Then, $P(x_1, \ldots, x_s \mid r_1, \ldots, r_s) = P(E \cap F)$. Let G be the event that none of the persons in $\mathscr{S}_1$ became infected by a person from $\mathscr{S}_0$. Then $P(E \cap F) = P(E \cap F \cap G) + P(E \cap F \cap \bar{G})$, where $\bar{G}$ is the complement of G. Obviously, $E \cap \bar{G}$ is the empty set, because $\bar{G}$ implies that at least one person in $\mathscr{S}_1$ was infected by a person in $\mathscr{S}_0$, while the E is the event that none of the members of $\mathscr{S}_0$ were infected. This yields

$$P(E \cap F) = P(E \cap F \cap G) = P(E \mid F \cap G)P(F \cap G). \tag{A2}$$

Now, $F \cap G$ is the event that all the members of $\mathscr{S}_1$ became infected either from the community or from other members of $\mathscr{S}_1$. Then, $P(F \cap G)$ is the probability that all the members in a group of k persons with risk levels $\{r_i \mid i \in \mathscr{S}_1\}$ became infected. The probability $P(F \cap G)$ is found recursively from set 3, below, for a household with k susceptibles $(k < s)$. The probability $P(E \mid F \cap G)$ is the probability that all the $s - k$ members of $\mathscr{S}_0$ escaped infection from the community as well as from the k members of $\mathscr{S}_1$. It is given by

$$P(E \mid F \cap G) = \prod_{i \in \mathscr{S}_0} B_{r_i} Q_{r_i}^{k}. \tag{A3}$$

Set 3: $x_1 = \cdots = x_s = 1$. The probability that all s household members are infected ($k = s$) is found by summing over the $2^s - 1$ other probabilities (for which $\sum x_i < s$) from sets 1 and 2 above and then subtracting that sum from 1.

$$P(\mathbf{1}_s \mid r_1, \ldots, r_s) = 1 - \sum_{x_1, \ldots, x_s: \sum x_i < s.} P(x_1, \ldots, x_s \mid r_1, \ldots, r_s). \tag{A4}$$

A special case of equation 1 seems to have been first noted by Bailey (21, equation 14.7, p. 248) for the Reed-Frost model. The derivation was extended for fixed length infectious periods by Ludwig (26) and Longini and Koopman (1) and for variable length infectious periods by Ball (27).

An example

To illustrate the recursive nature of equation 1, consider the following example: Suppose that $s = 3$, $R = 2$ and the objective is to find the probability of the infection array (1, 1, 0) given the risk factor array (1, 2, 2). From equations A2–A3,

$$P(1, 1, 0 \mid 1, 2, 2) = P(1, 1 \mid 1, 2)B_2Q_2^2. \tag{A5}$$

Thus, the term $P(1, 1 \mid 1, 2)$ must be found recursively from the case where $s = 2$, using equation A4, which yields

$$P(1, 1 \mid 1, 2) = 1 - P(0, 0 \mid 1, 2) - P(1, 0 \mid 1, 2) - P(0, 1 \mid 1, 2). \tag{A6}$$

From equations A1–A3,

$$P(0, 0 \mid 1, 2) = B_1B_2, \tag{A7}$$

$$P(1, 0 \mid 1, 2) = P(1 \mid 1)B_2Q_2, \tag{A8}$$

$$P(0, 1 \mid 1, 2) = P(1 \mid 2)B_1Q_1. \tag{A9}$$

The terms $P(1 \mid 1)$ and $P(1 \mid 2)$ in equations A8 and A9 are found recursively from the case where $s = 1$, using equation A4, which yields

$$P(1 \mid 1) = 1 - P(0 \mid 1) \quad \text{and} \quad P(1 \mid 2) = 1 - P(0 \mid 2). \tag{A10}$$

From equation A1,

$$P(0 \mid 1) = B_1 \quad \text{and} \quad P(0 \mid 2) = B_2. \tag{A11}$$

Substituting equations A6–A11 into equation A5 yields the final probability as

$$P(1, 1, 0 \mid 1, 2, 2) = [1 - B_1B_2 - (1 - B_1)B_2Q_2 - (1 - B_2)B_1Q_1]B_2Q_2^2.$$