AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 115, No 5

Printed in U.S.A.

ESTIMATING HOUSEHOLD AND COMMUNITY TRANSMISSION PARAMETERS FOR INFLUENZA

IRA M. LONGINI, JR.,[1] JAMES S. KOOPMAN,[1] ARNOLD S MONTO,[1] AND JOHN P. FOX[2]

Longini, I. M., Jr. (School of Public Health, University of Michigan, Ann Arbor, MI 48109), J. S. Koopman, A. S. Monto and J. P. Fox. Estimating household and community transmission parameters for influenza. *Am J Epidemiol* 1982; 115:736-51.

A maximum likelihood procedure is given for estimating houshold and community transmission parameters from observed influenza infection data. The estimator for the household transmission probability is an improvement over the classical secondary attack rate calculations because it factors out community-acquired infections from true secondary infections. The mathematical model used does not require the specification of infection onset times and, therefore, can be used with serologic data which detect asymptomatic infections. Infection data were derived by serology and virus isolation from the Tecumseh Respiratory Illness Study and the Seattle Flu Study for the years 1975–1979. Included were seasons of influenza B and influenza A subtypes H1N1 and H3N2. The transmission characteristics of influenza B and influenza A(H3N2) and A(H1N1) outbreaks during this period are compared. Influenza A(H1N1), A(H3N2) and influenza B are found to be in descending order both in terms of ease of spread in the household and intensity of the epidemic in the community. Children are found to be the main introducers of influenza into households. The degree of estimation error from the misclassification of infected and susceptible individuals is illustrated with a stochastic simulation model. This model simulates the expected number of detected infections at different levels of sensitivity and specificity for the serologic tests used. Other sources of estimation error, such as deviation from the model assumption of uniform community exposure and the possible presence of superspreaders, are also discussed.

biometry; influenza; serology

An understanding of the transmission dynamics and determinants of influenza in communities is necessary in order to formulate effective control strategies. Both stochastic simulation (1) and deterministic (2) models have been used to study the transmission and control by immunization of influenza A. These studies have been instrumental in identifying which segments of the community are most active in spreading virus. The effect of immunization of these groups on transmission and attack rates has been predicted by these models. Nevertheless, the basic problem of estimating parameters of transmission from household and

Received for publication June 8, 1981, and in final form September 29, 1981.

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

[1] Dept. of Epidemiology, School of Public Health, The University of Michigan, Ann Arbor, MI 48109. (Reprint requests to Dr Longini.)

[2] Dept. of Epidemiology, SC-36, School of Public Health and Community Medicine, University of Washington, Seattle, WA

This research was partially supported by the National Institutes of Health under National Research Service Award AI-6103-01 and National Institute of Allergy and Infectious Diseases, Development and Applications Branch Contract No. AI-62514.

community infection data has not been resolved. Traditionally, the secondary attack rate (SAR) has been used to measure household transmission. The classical SAR calculation has been shown to lead to serious error (3) by failing to separate true secondary household infections from community-acquired infections as well as from third and fourth generation infections. Consequently, the use of the SAR has come into question (1). A model has been developed to estimate the SAR as well as the probability that a household member will be infected from the community (4). The model not only provides parameter estimates for more elaborate dynamic models (1, 2), but also provides a better means of analyzing the social and environmental determinants of transmission when using serologic data. In this paper, we describe the development of the model, its robustness regarding basic assumptions, and its use to analyze transmission of different strains of influenza.

The data

Description. The data were taken from the Family Flu Study (1975–1979) (5) in Seattle and the Tecumseh Study of Respiratory Illness (6, 7). Details of study design and past data acquisition are given for the Seattle study in Fox and Hall (5) and for the Tecumseh study in Monto et al. (6) and Monto and Kioumehr (7). Detailed descriptions and analyses of the data from influenza seasons covered in this paper are forthcoming (Fox et al. and Monto et al., unpublished manuscripts). Only a brief summary of each study design is given below.

In the Seattle study, all households under study contain at least one child of five to 11 years of age. Virus isolation (nose-throat swabs) specimens are collected on a regular schedule from all family members at least biweekly (weekly in the influenza season) and at each onset of illness. All family members are bled every four months. Both hemagglutination-inhibition and complement-fixation tests are performed on the bloods.

In the Tecumseh study, the number of households on report represent 10 per cent of the community's population and households under study represent a cross-section of the population. Each household is then contacted weekly by telephone for illness symptom information and all illnesses are followed to their conclusion. Specimens for virus isolation are collected when a respiratory illness is reported within two days of onset. Blood specimens are collected every six months on a staggered basis, so that in any month, approximately one sixth of the population on report is having blood obtained. The hemagglutination-inhibition test is performed on the bloods.

For both studies, the influenza season or period of observation was confined to that period where virus isolation for a particular strain was most prevalent. Increased influenza-like morbidity was also used as a criterion for establishing the period of observation. Each period of observation was bracketed by pre- and post-season bleedings. The pre-season hemagglutination-inhibition titer was used to establish the susceptibility of each individual. Susceptible individuals were considered to have been infected during the course of the influenza season if there was a "significant" rise in antibody titer when comparing pre- and post-season titers. In addition, an individual was considered to have been infected if virus was isolated. The definition of what constitutes a significant rise was somewhat different for the Seattle (8) and Tecumseh (7) studies. These differences will not be discussed here. For the most part, a 4-fold rise was considered to be significant in both studies.

Table 1 shows the infection attack rates for individuals with different pre-season titer levels to influenza A(H3N2) (1977–1978) in Tecumseh. There was no age-related difference in the relationship

TABLE 1

Infection rates by pre-season hemagglutination-inhibition (HI) titer· influenza A(H3N2) 1977–1978 Tecumseh data

Pre-HI Titer	No. observed	Fraction infected
<1/8	478	0.203
1/8	111	0.144
1/16	72	0.028
1/32	61	0.082
1/64*	30	0.067
≥1/128	10	0

* All individuals with pre-HI titer ≤1/64 are considered to be susceptible, regardless of age.

between the infection rate and pre-season hemagglutination-inhibition titer level. Therefore, all individuals with a pre-season titer ≤1/64 were considered to be initially susceptible. Establishing susceptibility cutoffs is quite arbitrary and the effects of misclassifying individuals are discussed later. Table 2 gives a summary of the influenza seasons studied here with age-specific attack rates.

Distribution of infections in households. From the above information, the number of initial susceptibles that were infected during the course of the influenza season was calculated for each household. Only households that had complete serologic data were included. This information was used to build a frequency distribution of infections in households. Such a distribution is shown in table 3 for the influenza A(H3N2) season of 1977–1978, in Tecumseh. If there were fewer than 10 total households with a specified number of susceptibles, those households were not included. The total number of households used for the influenza B 1975–1976, A(H3N2) 1977–1978, and A(H1N1) 1978–1979 seasons in Seattle were 87, 159, and 93, respectively. The total number of households used in the influenza A(H3N2) (1977–1978) season in Tecumseh was 195. Table 3 shows that 44 households had only one initial susceptible and the single susceptible was infected in four of the 44 households. Also, 77 households had two initial susceptibles; in 12 households, one susceptible was infected, while two households had both susceptibles infected. The frequency distributions for the other influenza seasons are given under "Results."

These frequency distributions contain information about transmission. If there is a preponderance of households with only one member infected, the indication is

TABLE 3

Distribution of influenza A(H3N2) infections: 1977–1978 Tecumseh data

No infected	No of susceptibles*/household				
	1	2	3	4	5
0	40	63	18	17	3
1	4	12	10	5	4
2		2	3	3	2
3			2	3	2
4				1	1
5					0
Total	44	77	33	29	12

* All individuals with pre-season hemagglutination-inhibition titer ≤1/64 are considered to be susceptible, regardless of age.

TABLE 2

Influenza seasons and infection attack rates as determined by hemagglutination-inhibition titer rises and isolates

Strain	Location	Season	Attack rate by age (years)*			
			0–4	5–19	20+	Total
B	Seattle	1/15/76–3/15/76	0 083	0 292	0.070	0.172
A(H3N2)	Seattle	12/15/77–2/28/78	0.158	0 337	0.111	0.224
A(H3N2)	Tecumseh	12/11/77–2/11/78	0.340	0.277	0.101	0.160
A(H1N1)	Seattle	12/15/78–3/31/79	0 346	0.524	0.023	0 290

* Proportion of each group infected.

that there is little household transmission. If the number of households experiencing multiple infections increases, this is a reflection of increased household transmission. However, such a shift is superimposed over the distribution of infections arising from the community. A great deal of information concerning community transmission is contained in the comparison between households that had no infections and those experiencing at least one infection. The model described in the next section is used to estimate, in an efficient manner, the degree of household transmission and community-acquired infection from various frequency distributions.

The model

The model centers on the number of susceptible individuals within a particular household who become infected during the course of the period of observation. It is assumed that the probability of community acquisition of infection and household transmission is homogeneous across all households. Define B as the probability that a susceptible household member escapes being infected from the community during the period of observation. If one or more of the susceptible household members becomes infected from the community, the probability that other household members will be infected from within the household must be considered. Now define Q as the probability that a susceptible household member escapes infectious contact from a single infected household member during his entire infectious period.

Using the parameters B and Q, the probability that j of k initial susceptibles become infected in a particular household during the period of observation can be found. This is the probability Pr $(j|k)$, which will be designated as m_{jk} to simplify notation.

The probability density function, m_{jk}, is derived in the Appendix and has the recursive form

$$m_{jk} = \binom{k}{j} m_{jj} B^{(k-j)} Q^{j(k-j)}, j < k,$$

and

$$m_{kk} = 1 - \sum_{j=0}^{k-1} m_{jk}.$$

Figure 1 shows the shape of m_{jk} for households with five original susceptibles (i.e., $k = 5$) when $B = 0.75$, for different values of Q. Note that the mass shifts to the right as Q decreases, indicating increased household transmission. In Figure 1a, there is no spread of infection among household members (i.e., $Q = 1$), and the disease in question is presumably not "infectious." In this situation without household transmission, m_{jk} reduces to the probability density function of the binomial distribution; i.e.,

$$m_{jk} = \binom{k}{j} (1 - B)^j B^{k-j}, j \leq k.$$

If it is assumed that there is spread only within the household and there are initially ι infectives within the household, the probability density function becomes

$$m_{jk} = \binom{k}{j} m_{jj} Q^{(\iota+j)(k-j)}, j < k,$$

and $\iota + j$ is the final number of infectives in the household. This equation provides the final size distribution for the Reed-Frost model and for a more general infection model which does not have the restrictions on the length of the latent and infectious periods required by the Reed-Frost model.

Methods

Estimation of B and Q. The probabilities B and Q are estimated from data (like those presented in table 3), using maximum likelihood methods. The likelihood function is given in the Appendix, although a complete description of the method is given elsewhere (4). The maximum likelihood estimation procedure also provides the asymptotic variances of

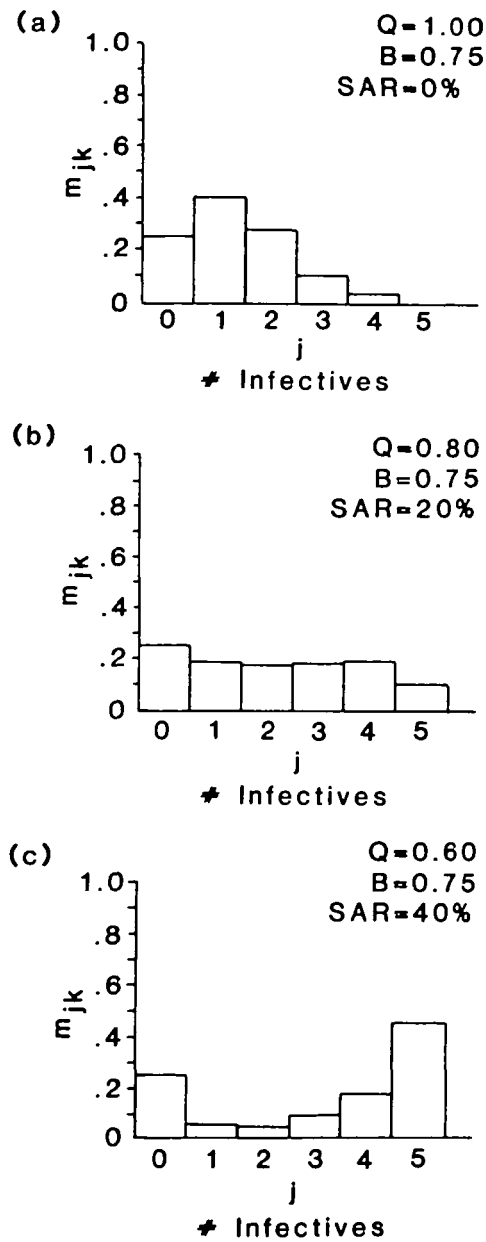

FIGURE 1. Probability density function m_{jk} for a household with five susceptible members ($k = 5$), when $B = 0.75$; (a) shows that when there is no secondary spread, i.e., $Q = 1$ or secondary attack rate (SAR) = 0, the probability density function (PDF) has the greatest mass around $j = 1$, indicating that most households will have only one infected individual; (b) shows that the PDF will be relatively flat when there is a moderate level of secondary spread; and (c) shows that a large number of households (nearly 50 per cent) will have all five members infected when secondary spread is high, i.e., $Q = 0.60$ or SAR = 40 per cent.

the estimators, which are approximately normal variates for large sample sizes.

Secondary attack rates and community infection probabilities. The estimators $\hat{Q}$ and $\hat{B}$ provide information on the degree of household transmission and infection in the community. The household secondary attack rate (SAR) has been used as a measure of household transmission. The classical definition of the SAR is given by Fox et al. (8). Use of this definition depends on knowing the onset times of household illnesses. However, serologic data that do not specify onset times can be used assuming that there is only a single primary infected individual and that all other infected individuals in the household are secondary to the single primary. Then, the classical definition of the SAR in each household is given by

$$\text{SAR} = \frac{\text{No. of infected household members} - 1}{\text{No. of initial susceptible household members} - 1} \times 100.$$

The above formula will result in third and fourth generation infections being counted as secondary infections. Therefore, the estimate of the SAR will be inflated over the SAR determined by the classical formula which uses infection onset times.

Use of the above formula or the classical formula using onset data will probably cause the estimate of the SAR to be inflated by apparent secondary infections that actually resulted from community exposure. There is, of course, no way of determining the true source of infection under normal field conditions. Therefore, community-acquired infections (other than primary infections) are often misclassified as secondary infections when using the classical estimation of the SAR. The magnitude of error increases as the risk of infection from the community increases (see table 2 in Kemper (3)). This problem and the problem of counting third and fourth generation infections are alleviated by using the estimated house-

hold transmission probability, $\hat{Q}$ from the model.

By definition, $1 - \hat{Q}$ is the probability that a single infective will make infectious contact with a household member during the course of the former's infectious period. Now, for example, assume that there is a single primary infective in a household that originally had five susceptibles. The expected number of new infectives due to contact with the primary infective is $4(1 - \hat{Q})$. It follows that the estimate of the household SAR is

$$\begin{aligned} \text{SAR} &= \left[4(1 - \hat{Q})/4\right] \times 100 \\ &= (1 - \hat{Q}) \times 100. \end{aligned}$$

This estimate of the SAR is not inflated by community-acquired infection or later generation infections and it will be referred to as the actual SAR, while the classical formula will be referred to as the apparent SAR.

For household studies, the introduction rate (i.e., the fraction of total households with at least one infected individual) has been used as a measure of community infection rates. Such a method will tend to underestimate the actual rate, since some infections following the introductory infection(s) will be due to the community. Therefore, the classical estimate for community infection rates is underestimated for the same reason that SARs are overestimated. This problem is alleviated by using $1 - \hat{B}$ as the measure of community risk to infection. By definition, $1 - \hat{B}$ is the probability that a household member will become infected from the community during the course of the influenza season. This probability will be referred to as the community probability of infection (CPI).

Model dependency on classification accuracy. While serologic classification of individuals as to their infection status is more accurate than using symptom data, a lack of sensitivity and possibly specificity of the serologic tests used could lead to some error. Since the hemagglutination-inhibition and complement fixation tests are not completely sensitive, some true infections will be misclassified as having remained susceptible during the course of the influenza season. This leads to underestimation of both the SAR and CPI. Lack of specificity, which is less of a problem, can have the opposite effect by misclassifying as infected those individuals who remained susceptible. Thus, the estimated SAR and CPI will tend to be inflated. When these two forms of misclassification operate in concert, they will have a joint effect, as demonstrated by the results of a simulation whose description follows.

A stochastic simulation of these errors was used to measure their effect on parameter estimation. A hypothetical population of 200 households, ranging in size from one to five members, was set up. Then, the probability density function for the hypothetical epidemic was generated with the *a priori* parameter values of $B = 0.800$, and $Q = 0.800$ — corresponding to CPI = 0.200 and actual SAR = 20.0. The sensitivity and specificity were set at some level and individuals were reclassified. For example, assume that the sensitivity and specificity are set at 0.700 and 0.900, respectively. Then, a random number between zero and one, inclusive, is generated for each infected individual. If that random number is less than or equal to 0.700, the individual remains classified as infected; if it is greater than 0.700, he is reclassified as being susceptible. Random numbers are also generated for uninfected individuals. If the random number is less than or equal to 0.900, the susceptible remains classified as uninfected; if it is greater than 0.900, he is reclassified as becoming infected. This procedure was carried out for every individual in each household and repeated 100 times. An average frequency distribution of infections in households was produced at various levels of sensitivity and specificity. The model was fit to the resulting frequency distribution and the

CPI and actual SAR were estimated. In this way, the effect of reduced sensitivity and specificity on parameter estimates could be measured.

Table 4 shows the estimated CPI and actual SAR from the simulated data at different levels of sensitivity and specificity. It is clear that the estimates for both the CPI and actual SAR decrease with decreasing sensitivity, as expected. Also note that at high levels of sensitivity, the actual SAR tends to first decrease and then increase as specificity decreases, while at lower levels of sensitivity, the actual SAR decreases monotonically with decreasing specificity. This is contrary to what is expected. Although the CPI does indeed increase with decreasing specificity, some of the increase is excessive due to the fact that the actual SAR is not inflated.

Results

Table 5 shows the fit of the model to the frequency distribution of infections in households that was given in table 3. The expected frequencies are calculated by substituting the maximum likelihood estimations $\hat{Q}$ and $\hat{B}$ back into the model. The fit is excellent, as indicated by the low chi-square goodness of fit statistic, which has a p value of 0.58. The CPI is estimated as 0.132, indicating that a household member had about 13 per cent chance of being infected from the community during the epidemic season. The household transmission probability was estimated as 0.147, which yields an actual SAR of 14.7 per cent. This indicates that a household member had about a 15 per cent chance of being infected by another household member during the course of the latter's infectious period. The apparent SAR was calculated at 25.9 per cent, which is inflated considerably above the actual SAR.

To evaluate the role that children play in the introduction of infection into households, the households were partitioned into two groups: those which had at least one member less than 18 years of age and those which had only adults. Tables 6 and 7 show the fit of the model to data from households with and without children, respectively. In both cases, the fit is quite good. Because households without children have so few susceptibles (usually one or two), there is little information available on secondary transmission. However, the estimated CPI for households with children is significantly higher ($p < 0.01$) than the CPI for households without children. By looking at the ratio of the CPIs, we conclude that the likelihood that an individual of any age from a household with a member less than 18 years of age will become infected from the community is almost two times greater than the likelihood that an adult from a household without children will be infected from the community. This suggests that the schools, day-care cen-

Table 4

Estimated actual SAR/CPI at different levels of sensitivity and specificity (stochastic simulation population where actual SAR = 20.0 and CPI = 0.200)*

Specificity	Sensitivity					
	1.00	0 95	0 90	0 85	0 80	0 75
1 00	20 0/0.200	18.2/0.196	16.8/0.191	15 4/0.184	14.2/0.177	13.0/0.170
0.95	18.0/0 240	16.6/0.232	15.2/0.224	13 8/0.216	12.7/0.207	11.1/0.199
0 90	17.0/0 284	15.6/0.274	13.7/0.263	12.5/0 253	11.1/0.241	9.9/0.231
0.85	18 1/0 321	15.5/0.311	13.3/0.302	11.7/0.292	10.0/0 281	8.9/0.268
0.80	18.7/0.365	15 8/0.355	13.3/0.341	11.3/0.328	9.8/0.314	8 3/0.301
0 75	20.0/0 407	16.5/0 394	13.8/0.383	11 3/0.368	9.4/0 354	7.9/0.388

* SAR, secondary attack rate; CPI, community probability of infection

TABLE 5
Observed and expected distribution of influenza A(H3N2) 1977–1978 Tecumseh data: all households

No infected	No of susceptibles* per household									
	1		2		3		4		5	
	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected
0	40	38.2	63	58.0	18	21.6	17	16.5	3	5 9
1	4	5.8	12	15 0	10	7 2	5	6 2	4	2.4
2			2	3 9	3	3.2	3	3 5	2	1.5
3					2	1 1	3	2 0	2	1.1
4							1	0 8	1	0.7
5									0	0 3
Total	44	44 0	77	76 9	33	33 1	29	29 0	12	11.9

* The criterion for classifying individuals as susceptible is: All individuals with pre-season hemagglutination-inhibition titer ≤1/64, regardless of age.
$Q = 0.853$; $B = 0\ 868$; $V(Q) = 0.0014$, $V(B) = 0.0003$; $\mathrm{Cov}(Q,B) = -0\ 0001$, $\chi^2(11) = 9.46$ (expected frequencies <1.0 are added); $p = 0.58$; actual SAR = 14.7; apparent SAR = 25.9.

ters, and other groups where children congregate play an important role in spreading the agent.

Tables 8 and 9 show the fit of the model to the data for the influenza B season of 1975–1976 and the influenza A(H1N1) season of 1978–1979, in Seattle. The frequency distribution for the influenza A(H3N2) is generally similar to the H3N2 agent in Tecumseh and is not shown. Table 10 gives a summary of all the agents under study. Each estimate is shown with its standard error. For the Seattle data, the CPIs for all three agents are found to be significantly different ($p < 0.01$). The conclusion is that influenza B caused a less intense epidemic than influenza A and that subtype H1N1 caused a more intense epidemic than H3N2. This is probably due to the large role played by schoolchildren in spreading H1N1 (7). In households, the actual SAR for influenza B is found to be significantly lower ($p < 0.05$) than the actual SAR for influenza A(H1N1). Although there is not a statistically significant difference in the actual SAR between influenza B and influenza A(H3N2) or influenza A(H3N2) and H1N1, there appears to be a tendency for H1N1 to spread more easily in the household than H3N2 and for H3N2 to spread more easily than influenza B.

The transmission parameters can also be compared for influenza A(H3N2) in Tecumseh and Seattle. When households with children in Tecumseh are compared to the Seattle households (all of which have children by the study design), there is a significant difference ($p < 0.01$) for the CPIs. This difference in the intensity of the epidemic caused by similar or identical agents could be due to a number of factors, especially age structure of the population. Also involved could be slight differences in the agent itself, prior antigenic experience of the population, social structure, study design, environmental factors, and pre-titer cutoffs used to establish susceptibility.

TABLE 6

Observed and expected distribution of influenza A(H3N2) 1977–1978 Tecumseh data: households with children 0–17 years of age

No infected	No. of susceptibles*/household					
	3		4		5	
	Observed	Expected	Observed	Expected	Observed	Expected
0	17	17.2	15	12.9	3	4.8
1	8	7.7	5	6 5	4	2.6
2	3	3.8	3	4.1	2	1.9
3	2	1.3	3	2.5	2	1 4
4			1	1.0	1	1.0
5					0	0.4
Total	30	30.0	27	27.0	12	12.1

* The criterion for classifying individuals as susceptible is: All individuals with pre-season hemagglutination-inhibition titer ≤1/64, regardless of age.

$\hat{Q} = 0.854$; $\hat{B} = 0.831$; $\text{Var}(\hat{Q}) = 0.0016$; $\text{Var}(\hat{B}) = 0.0007$; $\text{Cov}(\hat{Q}, \hat{B}) = -0\ 0003$; $\chi^2(9) = 3\ 43$ (expected frequencies <1.0 are added); $p = 0.90$; actual SAR = 14.6; apparent SAR = 28.6.

Table 10 also gives the daily household contact rate, $\hat{p}$, where $\hat{p}$ is defined as the daily probability that an infected household member will make infectious contact with a susceptible household member. The method for estimating $\hat{p}$ from the model parameter Q is described in the Appendix. The daily household contact rate is an important parameter in simulation models of influenza. In the influenza simulations of Elveback et al. (1), the value of $\hat{p}$ was set at 0.02 for simulations of the Asian (H2N2) and Hong Kong (H3N2) pandemic strains. Note that the estimated values of $\hat{p}$ from influenza A(H3N2) in Seattle and Tecumseh are considerably higher than 0.02. Although the estimated values of $\hat{p}$ are not identical to those in the simulations, they could serve as initial "guestimates" of parameters for future influenza simulations.

TABLE 7

Observed and expected distribution of influenza A(H3N2) 1977–1978 Tecumseh data: households without children 0–17 years of age

No. Infected	No of susceptibles*/household			
	1		2	
	Observed	Expected	Observed	Expected
0	40	40 2	61	60.8
1	4	3.8	11	11.1
2			1	1.0
Total	44	44 0	73	72.9

* The criterion for classifying individuals as susceptible is: All individuals with pre-season hemagglutination-inhibition titer ≤1/64, regardless of age.

$\hat{Q} = 0.960$; $\hat{B} = 0.913$; $\text{Var}(\hat{Q}) = 0.0070$; $\text{Var}(\hat{B}) = 0.0004$; $\text{Cov}(\hat{Q},\hat{B}) = -0.0002$; $\chi^2(1) = 0\ 01$ (expected frequencies <1 0 are added); $p = 0.92$; actual SAR = 4.0; apparent SAR = 8.3

DISCUSSION

The objective of any estimation procedure is to use the available information efficiently. The maximum likelihood procedure for estimating the SAR is both efficient and uninflated by community-acquired or third and fourth generation infections. It has a further advantage over classical methods of SAR estimation, since serologic data obtained from sequential serosurveys can be used. Use of such data is possible since the method requires no critical assumptions concerning timing of infections or the length of latent, incubation and infectious periods.

Use of the classical method for calculating SARs requires specifying the onset times of clinical infections. This requires that illness data be used, which means that noninfluenza illness may be counted as influenza cases and that asympto-

TABLE 8

Observed and expected distribution of influenza B 1975–1976 Seattle data

No infected	No of susceptibles*/household									
	1		2		3		4		5	
	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected
0	9	8.3	12	13.9	18	16.2	9	9 6	4	3.6
1	1	1.7	6	4 9	6	7.4	4	5.1	3	2.1
2			2	1 3	3	3 4	4	3.1	0	1.5
3					1	1 0	3	1.6	2	1 0
4							0	0 6	0	0.6
5									0	0.2
Total	10	10.0	20	20.1	28	28.0	20	20 0	9	9.0

* The criteria for classifying individuals as susceptible are: (a) all individuals <20 years of age regardless of pre-season hemagglutination-inhibition titer; (b) individuals ≥20 years of age with pre-season hemagglutination-inhibition titer ≤1/10.

$\hat{Q} = 0.873$; $\hat{B} = 0.833$, $\mathrm{Var}(\hat{Q}) = 0.0024$; $\mathrm{Var}(\hat{B}) = 0.0007$; $\mathrm{Cov}(\hat{Q},\hat{B}) = -0.0003$; $\chi^2(10) = 4.59$ (expected frequencies <1.0 are added); $p = 0.92$; actual SAR = 12.7%; apparent SAR = 25.9%.

matic infections, which can be quite important in transmission, will not be counted at all. Virus isolation data can be used to specify infection onset times for asymptomatic infections, but the cost of such a study is prohibitive. Even if accurate onset times can be established, the task of separating out chains of infection can be difficult since the infectious periods of successive generations of persons infected will tend to overlap. This problem is compounded when individuals are also acquiring infection from the community. The problems associated with using the classical estimation of the SAR have been discussed by others (1, 3, 9, 10). Use of the actual SAR eliminates a source of error by effectively separating household from community infection.

A second source of error is due to a lack of sensitivity and possibly specificity of the serologic tests used. The effects of this error source on the estimated CPI and SAR were demonstrated through simulated data under "Methods." The results of the simulation show that we can expect the underestimation of the CPI and actual SAR with decreased sensitivity and overestimation of only the CPI with decreased specificity. It is possible that the differences between the CPI and actual SAR seen among different influenza types and subtypes could be partially due to a lack of sensitivity. However, on the basis of the simulation results shown in table 4, it is doubtful that a lack of sensitivity accounts for the findings given.

The cutoffs in pre-titer, used to establish whether an individual is susceptible or immune prior to the epidemic season (see table 1), can lead to errors. If the cutoffs are too low, susceptible individuals may be misclassified as immune. This can lead to inflated estimates of the actual SAR and CPI by underestimating the population at risk to infection. Conversely, if the cutoffs are too high, underestimation may occur. The effect of this source of error on estimation was investi-

TABLE 9

Observed and expected distribution of influenza A(H1N1) 1978–1979 Seattle data

No infected	No. of susceptibles*/household					
	1		2		3	
	Observed	Expected	Observed	Expected	Observed	Expected
0	15	14.0	12	14.5	4	2.7
1	11	12.0	17	17.3	4	3.3
2			21	18 3	4	4.8
3					5	6 2
Total	26	26 0	50	50.1	17	17.0

* The criteria for classifying individuals as susceptible are: (a) individuals <20 years of age with pre-hemagglutination-inhibition titers ≤1/10; (b) all individuals ≥20 years of age are considered to be immune.

$\hat{Q} = 0.694$; $\hat{B} = 0.539$; $\mathrm{Var}(\hat{Q}) = 0\ 0076$; $\mathrm{Var}(\hat{B}) = 0.0020$, $\mathrm{Cov}(\hat{Q},\hat{B}) = -0.0015$, $\chi^2(4) = 2.22$ (expected frequencies <1.0 are added); $p = 0.695$; actual SAR = 30.6%; apparent SAR = 54.7%.

gated by trying different cutoff points for various data sets. The magnitude of error was found to be quite small when reasonable cutoffs were used. To remove this source of error, a function involving relative susceptibility (1) should be used. The relative susceptibility would then be a decreasing function of the pre-hemagglutination-inhibition titer level. An analytical development of models involving varying susceptibility, which is given by Ludwig (11), is quite complex. A more direct method is simply to adjust the data for the relative susceptibility of each individual.

A fourth source of error involves the assumption that the probability of community acquisition of infection is homogeneous across all households. Certainly this assumption is not satisfied in reality. This nonhomogeneity will cause the estimate of the actual SAR to be overestimated and the average CPI to be underestimated. This source of error could be partially controlled by partitioning households into more homogeneous groups. This is illustrated for influenza A(H3N2) (1977–1978) in Tecumseh (see table 10). The magnitude of error introduced by failure to meet the assumption of homogeneity needs to be explored, although we doubt that this source of error could have accounted for the observed parameter differences among the epidemics compared in this paper. A useful extension of the model would be to estimate separate CPIs for different age groups. This would require the inclusion of separate Bs in the probability density function for each age group.

The presence of superspreaders could also account for nonhomogeneity of community-acquired infection. Elveback et al. (1) demonstrated that the presence of superspreaders would be hard to detect from attack rate data, while Kemper (12) demonstrated analytically that their presence could not be detected from incidence and prevalence data.

The methods presented in this paper, along with the results of the paper by Kemper, should help to establish a sound theoretical basis for the use and interpretation of the SAR. Serologic methods can be used to include all infections rather than a fraction of the infections that is available from symptom data. The actual SAR can then be estimated by the model which distinguishes household from community-acquired infections. The use of separate transmission parameter estimates at the household and community levels overcomes disadvantages in the use of crude attack rates to relate to transmission factors. Namely, when using crude attack rates, there may be so many

TABLE 10

Summary of influenza agents studied from Tecumseh and Seattle, 1975–1979

Strain of influenza	Source	Sample size and description	Actual SAR* $(1 - \hat{Q}) \times 100$	Apparent SAR	CPI* $1 - \hat{B}$	Daily household contact probability $\hat{p}$†
Influenza B						
1975–1976	Seattle	87 households with children	12 7 ± 4.9	25 9	0.167 ± 0 026	0 033 ± 0.014
Influenza A						
A(H3N2) 1977–1979	Seattle	159 households with children	20.6 ± 4 1	37.2	0.259 ± 0.026	0.056 ± 0.012
A(H3N2) 1977–1978	Tecumseh	195 households, all ages	14.7 ± 3 7	25.9	0.132 ± 0.017	0.039 ± 0.011
		69 households with children	14.6 ± 4.0	28 6	0 169 ± 0.027	0.039 ± 0 011
		117 households without children‡			0.087 ± 0.021	
A(H1N1) 1978–1979	Seattle	93 households with children	30.6 ± 8 7	54.7	0.461 ± 0.017	0.087 ± 0.029

* SAR, secondary attack rate; CPI, community probability of infection.

† See the Appendix for the derivation of $\hat{p}$.

‡ The SAR information is not given because the SAR estimator is inefficient for households with only one or two members.

transmission factors causing infection that the relationship of any specific factor to infection may be obscured. This might be especially true for factors acting at the community level where the factor causes a small proportion of the infections directly but could be important in initiating chains of infection (13). Such factors would indeed cause a greater change in the CPI than in the crude attack rate. Similarly, there may be factors acting in the household which cause only a small proportion of the infections overall because of frequent community sources of infection. Such factors will again cause a greater change in the actual SAR than in the crude attack rate.

Together, the actual SAR and the CPI can be used to classify the transmission characteristics of influenza and other infectious agents. Certainly, the actual SAR and the CPI should relate to the mode of transmission of different infectious agents. One would expect airborne agents to have a relatively high CPI in relation to the actual SAR. In contrast, agents that are spread in a more direct fashion would tend to have a relatively low CPI in relation to the actual SAR. Even if one mode of transmission caused more deviation from the assumption of uniform probability of community acquisition of infection than another mode caused, the magnitude of such a deviation in itself could be useful for classification purposes. Therefore, the methodology presented here could be very helpful in classifying and exploring different modes of transmission among various infectious agents.

Besides its use in the study of the determinants of transmission, the methodology presented here could be used to provide parameter estimates for models of public health policy decision making (1, 2). The model presented here has utility in both scientific description and in predictions of public health consequences.

References

1. Elveback LR, Fox JP, Ackerman E, et al. An influenza simulation model for immunization studies. Am J Epidemiol 1976;103:152-65.
2. Longini IM, Ackerman E, Elveback LR. An optimization model for influenza A epidemics. Math Biosci 1977;38:141-57
3. Kemper JT. Error sources in the evaluation of secondary attack rates. Am J Epidemiol 1981;112:457-64
4. Longini IM, Koopman JS. Household and community transmission parameters from final distributions of infections in households Biometrics (in press)
5. Fox JP, Hall CE. Viruses in families. Littleton, MA: PSG, 1980.
6. Monto AS, Napier JA, Metzner HL The Tecumseh study of respiratory illness 1 Plan of study and observations on syndromes of acute respiratory disease. Am J Epidemiol 1971;94 269-79.
7. 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.
8. Fox JP, Hall CE, Elveback LR. Epidemiology: man and disease London. Collier-MacMillan, 1970:149.
9. Fox JP Family-based epidemiological studies. the second Wade Hampton Frost lecture Am J Epidemiol 1974;99:165-79.
10. Fox JP, Cooney MK, Hall CE. The Seattle virus watch. V Epidemiologic observations of rhinovirus infections, 1965-1969, in families with young children Am J Epidemiol 1975;101 122-43.
11. Ludwig D. Final size distributions for epidemics. Math Biosci 1975;23:33-46.
12. Kemper J. On the identification of superspreaders for infectious disease Math Biosci 1980;48: 111-27
13. Koopman JS, Guzman N, Henao O, et al. A diarrheal disease surveillance system in Cali, Colombia: theoretical basis and methods. Bull Pan Am Health Organ 1978,12:323-34.
14. Kendall MG, Stuart A The advanced theory of statistics, Vol I, 2nd ed London: Griffin, 1963:232.

Appendix

Derivation of the probability density function m_{jk}. Assume that household members mix randomly among themselves and that the probability that a household member is infected in the community is not affected by the number of infected members in his

household. Also, assume that all households under consideration are free of infected members at the beginning and end of the period of observation.

Using similar ideas to those of Ludwig (11) concerning final value distributions, the values of m_{jk} are derived as follows: When $k = 1$, it follows from the above assumptions that

$$m_{01} = B \text{ and } m_{11} = 1 - B.$$

Since there is random mixing, when $k = 2$, we have

$$m_{02} = B^2.$$

When considering m_{12}, there are two ways this can occur. Either the first susceptible individual becomes infected with probability $1 - B$, and the second escapes infection (from both the infective in the household, with probability Q, and in the community at large), or the first susceptible individual escapes and the second does not. It follows that

$$m_{12} = 2(1 - B)BQ = 2m_{11}BQ.$$

Deriving m_{22} follows a similar argument to the one above, so that

$$m_{22} = 2(1 - B)(1 - Q)B + (1 - B)^2 = 1 - m_{02} - m_{12},$$

as expected, since the probabilities must sum to one.

In general, there are $\binom{k}{j}$ ways to get j finally infected from k originally susceptible individuals. The $k - j$ susceptible individuals who escape infection must avoid having infectious contact with the j infective individuals in their household and from the community. The general expression for m_{jk} is

$$m_{jk} = \binom{k}{j} m_{jj} B^{(k-j)} Q^{j(k-j)}, j < k,$$

and

$$m_{kk} = 1 - \sum_{j=0}^{k-1} m_{jk}.$$

Maximum likelihood estimation of parameters. The parameters of interest, Q and B, are estimated using maximum likelihood methods. If there are n households in total, then let a_{jk} ($k = 1, 2, \ldots, K$ and $j = 0, 1, \ldots, k$) be the number of households observed that had j of k initial susceptibles infected during the period of observation. Then, the likelihood function (L) is

$$L(Q,B) = \prod_{k=1}^{K} \prod_{j=0}^{k} [m_{jk}]^{a_{jk}},$$

where m_{jk} is given by the recursive probability density function above. Substituting for m_{jk} and taking the log of the likelihood function yields

$$\ln L = C + \sum_{k} \sum_{j} a_{jk} [\ln m_{jj} + (k - j) \ln B + j(k - j) \ln Q],$$

where C is a constant.

The maximum likelihood estimators $\hat{Q}$ and $\hat{B}$ are solutions of

$$0 = \left.\frac{\partial \ln L}{\partial Q}\right|_{\hat{Q},\hat{B}} = \sum_{k=1}^{K}\sum_{j=0}^{k} a_{jk}\left[\frac{1}{m_{jj}}\left(\frac{\partial m_{jj}}{\partial Q}\right) + \frac{j(k-j)}{Q}\right]$$

$$0 = \left.\frac{\partial \ln L}{\partial B}\right|_{\hat{Q},\hat{B}} = \sum_{k=1}^{K}\sum_{j=0}^{k} a_{jk}\left[\frac{1}{m_{jj}}\left(\frac{\partial m_{jj}}{\partial B}\right) + \frac{(k-j)}{B}\right]$$

These are solved iteratively using the method of scoring. The elements of the information matrix are given by the expected values, with respect to a_{jk}, of the second partials. The information matrix, as well as further details on estimation, is given elsewhere (4).

Calculation of apparent SAR. In this calculation, it is assumed that there is always one primary infective (i.e., no co-primaries) and that all other susceptibles in the household are exposed to that primary. Then the apparent SAR is calculated as

$$\text{SAR} = \frac{\sum_{k=2}^{K}\sum_{j=2}^{k}(j-1)\,a_{jk}}{\sum_{k=2}^{K}(k-1)\sum_{j=1}^{k} a_{jk}}$$

Calculation of daily household contact probability and its variance. Assume that an individual who is infected in time period t_0 will pass through a series of stages at time period $t_1, t_2, \ldots,$ until he becomes immune. Define p_t as the probability that an infective who was infected in time period $t = t_0$ will make infectious contact in the household with another individual in time period t. Then, $\{p_t\}$ describes the pattern of infectiousness over time. The structure of $\{p_t\}$ is

$$\begin{aligned} p_t &= 0 \text{ when } t_0 \leq t \leq t_l \text{ latent period,} \\ p_t &> 0 \text{ when } t_{l+1} \leq t \leq t_m \text{ infectious period,} \\ p_t &= 0 \text{ when } t_{m+1} \leq t \leq t_n \text{ immune period.} \end{aligned} \qquad \text{1a}$$

Let $q_t = 1 - p_t$ be defined as the probability of escaping infectious contact. Then, the probability Q that the susceptible individual escapes infectious contact from the infective during his entire period of infectiousness is

$$\hat{Q} = \prod_{t=t_{l+1}}^{t_m} q_t, \qquad \text{2a}$$

where $\hat{Q}$ is the maximum likelihood estimator for Q.

For influenza, the average length of the latent and infectious periods are about two and four days, respectively. Therefore, in equation 1a, $l = 2$ and $m = 6$. If it is assumed that $p_t = p$ for $t = 3, \ldots, 6$, then application of equation 2a yields

$$\hat{Q} = \hat{q}^4 \text{ and } \hat{q} = \hat{Q}^{1/4} \qquad \text{3a}$$

Since the variance of $\hat{Q}$ is known, the variance of $\hat{q}$ is a function of a known variance. The asymptotic variance of $\hat{q}$ is

$$\text{Var}(f(\hat{Q})) = \left(\frac{df}{d\theta}\right)^2 \text{Var}(\hat{Q}),$$

where

$$f(\hat{Q}) = \hat{Q}^{1/4} \equiv \hat{q} \text{ and } \theta = \hat{Q}.$$

(see reference 14). Therefore,

$$\text{Var}(\hat{q}) = (1/4)^2 \hat{Q}^{-3/2} \text{Var}(\hat{Q}),$$

and

$$\text{Var}(\hat{p}) = \text{Var}(\hat{q}),$$

where

$$\hat{p} = 1 - \hat{q}.$$

Example: Influenza A (H1N1) 1978–1979 Seattle data in table 10.

$$\hat{Q} = 0.694 \text{ and } \text{Var}(\hat{Q}) = 0.0076$$

From equation 3a, $\hat{q} = 0.913$ and $\hat{p} = 0.087$

$$\text{Var}(\hat{p}) = (1/4)^2(0.694)^{-3/2}(0.0076) = 0.00082$$

and the standard error of $\hat{p}$ is 0.029.