ANALYSIS OF INFECTIOUS DISEASE DATA FROM A SAMPLE OF HOUSEHOLDS

Niels G. Becker
La Trobe University, Australia.

Summary

Observations are made on the individuals from a sample of households at two points in time. At the first time point it is determined who is susceptible to a certain infectious disease. At the second time point it is determined which of the susceptibles have been infected. A method for the analysis of such data is derived with the aid of martingale theory and suitable epidemic model assumptions. In contrast to a previously proposed analysis the present approach requires only a modest amount of manual calculation and is based on less restrictive assumptions. An application to influenza data is given.

1. Introduction Consider the spread of an infectious disease through a community of households. Suppose the disease is transmitted primarily by sufficiently close contacts between susceptible and infectious individuals, and that such contacts may take place between any pair of individuals.

1.1. Data on Sizes of Outbreaks in Households. Each of the individuals from a random sample of households is tested at time 0 and again at time τ. At time 0 it is determined who is susceptible to the infectious disease and at time τ it is determined which of the susceptibles has been infected since time 0. The time interval $(0, \tau)$ might be the time interval during which the epidemic passes through this community, or the 'epidemic season' of a disease which is endemic in the community. It is important to consider this type of infectious disease data, because it is relatively easy to obtain and is often reliable. When based on tests of sera one tends to obtain more objective data and can hope to cope with the presence of subclinical infections.

It has been the practice to analyse such infectious disease data by assuming that outbreaks within households essentially evolve independently of each other. However, it seems safer to base the analysis on a model which permits both between and within household infection. Such an analysis is described by Becker & Hopper (1983) for a community in which every household is kept under observation. Complete observation is impractical for large communities. Haber et al. (1988) propose an analysis which can be applied to data from a random sample of households from a larger community. Here we demonstrate that the ap-

proach of Becker and Hopper applies under the assumptions used by Haber et al. There results an analysis which involves only simple calculations which are easily performed manually. Furthermore, the assumptions are not quite as restrictive as those of the model used by Haber et al.

The proposed analysis is derived using martingale theory for counting processes. For a review of the requisite results from that theory see Andersen & Borgan (1985). We are able to give explicit expressions for estimates and standard errors. Central limit theorems also exist, see Andersen & Borgan (1985), which permits hypothesis testing and the construction of confidence intervals.

1.2. Model Assumptions and the Basic Martingales. Consider a random sample of n households. Label these 1,2,...,n. Let $S_h(t)$ denote the number of individuals from household h who are still susceptible at time t. Write s_h for $S_h(0)$. Of the $N_h(t)=s_h-S_h(t)$ individuals infected by time t there are $I_h(t)$ infectious individuals at time t.

The infection intensity acting on each of the $S_h(t)$ susceptibles of household h at time t consists of two components. One arises from within the household and is given by $\beta I_h(t)$, while the other arises from the remaining community and is denoted by λ_t. Note that λ_t depends on the total number of infectives in the community at time t. This makes it more realistic to view λ_t as a random process. Randomness in λ_t can be accommodated within the present framework, but is excluded under the assumptions used by Haber et al. (1988).

Let F_t denote the history of the infection process up to time t. Then

$$M_{1h}(t) = N_h(t) - \beta\int_0^t I_h(x)\,S_h(x)\,dx - \int_0^t \lambda_x S_h(x)\,dx \quad h=1,2,...,n$$

are orthogonal zero-mean martingales (ZMMs) with respect to the filtration $\{F_t\}$.

Further characteristics of the disease need to be specified. No specific assumptions are made about the latent period. We assume that the infectious period has a random duration and thereafter the infected individual becomes immune from further infection for the remaining duration of the observation period. Individuals who end their infectious period are called removals. Let $R_h(t)$ be the number of removals in household h at time t. In order to simplify the derivation we assume that the duration of the infectious period is exponential with rate parameter γ. This assumption is not crucial to the derivation and when this as-

sumption is relaxed, see Becker (1979, 1981), one arrives at the same estimates. However, expressions for standard errors need to be modified slightly. The processes specified by

$$M_{2h}(t) = R_h(t) - \gamma \int_0^t I_h(x)\,dx \qquad h=1,2,...,n$$

are orthogonal ZMMs, with respect to $\{F_t\}$, under the present assumptions.

The processes $M_{1h}(t)$ and $M_{2h}(t)$ h=1,2,...,n are the basic martingales used to derive the method of inference.

We will be integrating step functions with respect to ZMMs on several occasions. The left continuous modification of these step functions will be used on each occasion, without this being indicated explicitly by the notation. This is done to ensure that the integration is over a predictable process.

2. Community Acquired Infection. The risk of being infected from outside the household over the interval $(0, \tau)$ is determined by $\Lambda_\tau = \int_0^\tau \lambda_x dx$, the integrated global infection intensity. Consider the estimation of this quantity.

2.1. The Integrated Global Intensity Function. For each h define a counting process K_h by

$$K_h(t) = 1\ \{N_h(t) > 0\},$$

which simply indicates whether household h has been infected by time t. Note that $K_h(t)$ can be expressed as $\int_0^t 1\{N_h(x) = 0\}\, dN_h(x)$. We use this observation to construct the ZMMs

$$M_{1h}^{\dagger}(t) = \int_0^t 1\{N_h(x) = 0\}\, dM_{1h}(x)$$

$$= K_h(t) - \int_0^t \lambda_x s_h \{1 - K_h(x)\}\, dx \qquad h=1,2,...,n.$$

The choice of $1\{N_h(x)=0\}$ as integrand is made so as to eliminate β from the expression. Its use restricts attention to the household only as long as it contains no infected individuals, which ensures that $1\{N_h(x)=0\}\ I_h(x)=0$.

By summing the $M_{1h}^{\dagger}$ over all households one obtains the ZMM

$$\kappa(t) = K_t - \int_0^t \lambda_x \left\{ s - \sum_h s_h K_h(x) \right\} dx,$$

where $s = \sum_h s_h$ and $K_t = \sum_h K_h(t)$, the total number of households infected by time t. A further integration leads to the ZMM

$$\kappa^*(t) = \int_0^t \frac{1\{K_x < n\}}{s - \sum_h s_h K_h(x)} d\kappa(x)$$

$$= \int_0^t \frac{dK_x}{s - \sum_h s_h K_h(x)} - \int_0^t \lambda_x 1\{K_x < n\} dx.$$

Equating this ZMM, evaluated at τ, to its mean indicates that

$$\hat{\Lambda}_\tau = \int_0^\tau \frac{dK_x}{s - \sum_h s_h K_h(x)} \tag{2.1}$$

is unbiased for $\int_0^\tau \lambda_x 1\{K_x < n\} dx$. The latter quantity is $\Lambda_\tau = \int_0^\tau \lambda_x dx$ unless every household contains at least one infected individual at time τ.

The bias $E\left(\hat{\Lambda}_\tau - \Lambda_\tau\right)$ is

$$E\int_0^\tau \lambda_x 1\{K_x = n\} dx = \int_0^\tau \lambda_x P(K_x = n) dx$$

$$= \int_0^\tau \lambda_x \prod_{h=1}^{n} (1 - e^{-s_h \Lambda_x}) dx,$$

which is a rapidly decreasing function of n. The bias is negligible in applications, where one should aim to have at least 50 households.

For the purpose of providing a standard error we need to consider the variance of $\kappa^*(t)$. From standard results in martingale theory, see Andersen & Borgan (1985, equations 3.15, 3.18), we find

$$V[\kappa^*(t)] = E\int_0^t \left[\frac{dK_x}{\left\{s-\sum_h s_h K_h(x)\right\}^2}\right]. \tag{2.2}$$

Households of Equal Size. Suppose all households are of the same size, that is $s_h=m$ $h=1,2,\ldots,n$. Then

$$\hat{\Lambda}_\tau = \frac{1}{m}\int_0^\tau \frac{dK_x}{n-K_x} = \frac{1}{m}\sum_{j=0}^{K_\tau-1}\frac{1}{n-j}$$

$$\cong \frac{1}{m} ln\left(\frac{n+\frac{1}{2}}{n-K_\tau+\frac{1}{2}}\right) \tag{2.3}$$

Also

$$\int_0^\tau \frac{dK_x}{\left\{s-\sum_h s_h K_h(x)\right\}^2} = \frac{1}{m^2}\int_0^\tau \frac{dK_x}{(n-K_x)^2} = \frac{1}{m^2}\sum_{j=0}^{K_\tau-1}(n-j)^{-2},$$

so that, in view of (2.2), we obtain

$$\text{s.e.}(\hat{\Lambda}_\tau) \cong \frac{1}{m}\left[\frac{1}{n-K_\tau+\frac{1}{2}} - \frac{1}{n+\frac{1}{2}}\right]^{\frac{1}{2}}. \tag{2.4}$$

Households of Varying Sizes. Often data are available from households of varying sizes. Strictly speaking, one is not able to compute $\hat{\Lambda}_\tau$ as given by (2.1), because the order in which the households are infected is not known. One way of overcoming this problem is to obtain a separate estimate of Λ_τ from each group of households of a given size. These estimates can then be pooled by taking their weighted average, with weights proportional to the inverse of their estimated variances. This should work well when the numbers in each group is moderately large. With small numbers there is a risk that a poor estimate receives a large weight.

An alternative approach is based on the observation that the value of (2.1) is much the same for all likely orderings of household infections. A good approximation to (2.1) is provided by calculating

m_a = (mean size of affected households)

and replacing $s - \sum_h s_h K_h(x)$ by $s - \sum_h m_a K_h(x)$. In other words, use

$$\hat{\Lambda}_\tau = \frac{1}{m_a}\int_0^\tau \frac{dK_x}{n^* - K_x} = \frac{1}{m_a}\sum_{j=1}^{K_\tau - 1}\frac{1}{n^* - j}$$

$$\cong \frac{1}{m_a} ln\left(\frac{n^* + \frac{1}{2}}{n^* - K_\tau + \frac{1}{2}}\right), \tag{2.5}$$

where $n^* = s/m_a$.

By a similar argument

$$\text{s.e. } (\hat{\Lambda}_\tau) \cong \frac{1}{m_a}\left[\frac{1}{n^* - K_\tau + \frac{1}{2}} - \frac{1}{n^* + \frac{1}{2}}\right]^{1/2}. \tag{2.6}$$

The expressions (2.5) and (2.6) agree exactly with (2.3) and (2.4), respectively, when all households are of the same size.

2.2. The Probability of Infection from Global Sources. The quantity $P_\tau = 1 - e^{-\Lambda_\tau}$ is of interest because of its meaningful epidemiological interpretation. It is the probability that an individual makes at least one infectious contact during $(0, \tau)$ with an infective from another household.

From data on outbreaks in n households of size m one can estimate P_τ by

$$\hat{P}_\tau = 1 - e^{-\hat{\Lambda}_\tau} \cong 1 - exp\left\{-\frac{1}{m} ln\left(\frac{n}{n - K_\tau}\right)\right\}$$

$$= 1 - \left(1 - \frac{K_\tau}{n}\right)^{1/m}.$$

The approximation holds when n is large. Using the delta method, see Rao (1973, section 6a.2), one obtains

$$\text{s.e. } (\hat{P}_\tau) = e^{-\hat{\Lambda}_\tau}\text{s.e. } (\hat{\Lambda}_\tau)$$

$$\cong e^{-\hat{\Lambda}_\tau}\frac{1}{m}\left[\frac{n}{n-K_\tau}-\frac{1}{n}\right]^{\frac{1}{2}}$$

$$=\frac{1}{m\sqrt{n}}\left(1-\frac{K_\tau}{n}\right)^{\frac{1}{m}-\frac{1}{2}}\left(\frac{K_\tau}{n}\right)^{\frac{1}{2}}.$$

These final expressions for $\hat{P}_\tau$ and s.e. $(\hat{P}_\tau)$ can also be arrived at by starting from the observation that, under the present model formulation, K_τ has a binomial distribution with index n and success probability parameter $e^{-m\Lambda_\tau}$. It is encouraging that the martingale approach leads to essentially the same estimates as the classical approach.

For households of varying sizes one simply replaces m and n by m_a and n*,respectively.

The estimate of Λ_τ given by this approach is essentially unbiased. By Jens en's inequality it follows that $1-e^{-\hat{\Lambda}_\tau}$ will tend to underestimate $P_\tau = 1-e^{-\Lambda_\tau}$. However, it shares with the maximum likelihood estimate the property that it is asymptotically unbiased.

3. The Within - Households Infection Potential. There is no continuous observation over time and so one is not able to estimate the rate parameters β and γ. We turn instead to the estimation of the infection potential $\theta=\beta/\gamma$. The interpretation of θ as an infection potential stems from the observation that it is of the form

(infection rate) x (mean duration of the infectious period).

Note that households of size 1 contain no information about θ. The discussion given below is therefore intended only for households having at least two susceptibles initially.

Consider again the basic martingales $M_{1h}(t)$ h=1,2,...,n. Let

$$H_{1h}(t) = 1\{K_t < n\}\ 1\{S_h(t) > 0\}/S_h(t)$$

and

$$H_{2h}(t) = 1\{K_t < n\}\ 1\{N_h(t) = 0\}\ (n - L_t)/\{s - \sum_h s_h K_h(t)\},$$

where L_t is the number of households with every individual infected by time t,

so that $\sum_h 1\{S_h(t) > 0\} = n - L_t$.

Now construct the two ZMMs

$$M_1^*(t) = \sum_h \int_0^t H_{1h}(x)\, dM_{1h}(x)$$

$$= \sum_h \int_0^t \frac{1\{K_x < n\}}{S_h(x)} dN_h(x)$$

$$-\beta \sum_h \int_0^t 1\{K_x < n\}\, 1\{S_h(x) > 0\}\, I_h(x)\, dx$$

$$-\int_0^t 1\{K_x < n\}\, \lambda_x (n - L_x)\, dx$$

and

$$M_1^{\wp}(t) = \sum_h \int_0^t H_{2h}(x)\, dM_{1h}(x)$$

$$= \int_0^t \frac{n - L_x}{s - \sum_h s_h K_h(x)} dK_x - \int_0^t 1\{K_x < n\}\, \lambda_x (n - L_x)\, dx.$$

Note that the choices of H_{1h} and H_{2h} have made the terms involving λ_x the same in these two ZMMs. The quantity λ_x is therefore eliminated by taking the difference $M_1^*(t) - M_1^{\wp}(t)$. However, there is one other term which needs attention. The term containing β contains the unobservable $I_h(x)$ quantities. To eliminate these unobservable terms we use, for the first time, the ZMMs $M_{2h}(t)$ h=1,2,...,n.

Construct the ZMM

$$M_2^*(t) = \sum_h \int_0^t \theta 1\{K_x < n\}\, 1\{S_h(x) > 0\}\, dM_{2h}(x).$$

We finally obtain the ZMM

$$M_1^*(t) - M_1^{\wp}(t) - M_2^*(t)$$

$$= \sum_h \int_0^t \frac{1\{K_x < n\}}{S_h(x)} dN_h(x) - \int_0^t \frac{n - L_x}{s - \sum_h s_h K_h(x)} dK_x$$

$$-\theta \sum_h \int_0^t 1\{K_x < n\}\, 1\{S_h(x) > 0\}\, dR_h(x) .$$

$$= A_t - B_t - \theta C_t, \text{ say.}$$

By evaluating this ZMM at τ and equating it to its mean we obtain the estimate

$$\hat{\theta} = (A_\tau - B_\tau)/C_\tau \quad .$$

It is not always possible to assign exact values to A_τ, B_τ and C_τ, but there are situations when one can.

Suppose $K_\tau < n$, which means that not every household becomes infected. Then

$$A_\tau = \sum_{h=1}^{n} \sum_{j=0}^{N_h(\tau)-1} \frac{1}{s_h - j}.$$

For B_τ and C_τ we substitute

$$n\hat{\Lambda}_\tau \cong \frac{n}{m_a} ln\left[\frac{n^* + 1/2}{(n^* - K_\tau + 1/2)}\right] \text{ and } R_\tau - 2L_\tau = \sum_{h=1}^{n} R_h(\tau) - 2L_\tau,$$

respectively. These are the exact values for B_τ and C_τ whenever no household is totally infected, that is when $L_\tau = 0$. Otherwise these values are likely to be larger than they should be and their use will tend to underestimate θ. For infectious diseases of low or moderate infectiousness L_τ will tend to be small, compared with n, and these values for B_τ and C_τ will be good approximations. The correction term $2L_\tau$ used for C_τ is somewhat arbitrary. For a household in which everyone gets infected, the term $\int_0^\tau 1\{K_x < n\}\, 1\{S_h(x) > 0\}\, dR_h(x)$ is essentially the number of removals at the time when the last susceptible is infected. This unobserved quantity takes values in $\{0,1,...,R_h(\tau)-1\}$. However, its value can be at most $R_h(\tau)-2$ when the last infection arises from a within-household contact. We have, somewhat arbitrarily, proposed the value $R_h(\tau)-2$.

We now obtain an expression for the standard error of $\hat{\theta}$. Using the orthogo-

nality of the ZMMs M_{1h} and M_{2h} h=1,2,...,n, we find

$$V[M_1^*(t) - M_1^{\wp}(t) - M_2^*(t)]$$

$$= V[M_1^*(t)] + V[M_1^{\wp}(t)] + V[M_2^*(t)]$$

$$-2\sum_h Cov\left[\int_0^t H_{1h}(x)\,dM_{1h}(x),\ \int_0^t H_{2h}(x)\,dM_{1h}(x)\right].$$

Now use known results from martingale theory, see Andersen & Borgan (1985, equations 3.15, 3.17 and 3.18), to find

$$V[M_1^*(t) - M_1^{\wp}(t) - M_2^*(t)]$$

$$= E\left[\sum_h \int_0^t H_{1h}^2(x)\,dN_h(x)\right] + E\left[\sum_h \int_0^t H_{2h}^2(x)\,dN_h(x)\right]$$

$$+\theta^2 E\left[\sum_h \int_0^t 1\{K_x < n\}\,1\{S_h(x) > 0\}\,dR_h(x)\right]$$

$$-2E\left[\sum_h \int_0^t H_{1h}(x)\,H_{2h}(x)\,dN_h(x)\right].$$

We need to estimate each of the four terms on the right hand side, with time t=τ. Unbiased estimates of these terms are given respectively, by

$$\sum_h \int_0^\tau \frac{1\{K_x < n\}}{S_h^2(x)}\,dN_h(x) \cong \sum_{h=1}^{n} \sum_{j=0}^{N_h(\tau)-1} \frac{1}{(s_{h-j})^2},$$

$$\int_0^\tau \frac{(n-L_x)^2}{\left\{s - \sum_h s_h K_h(x)\right\}^2}\,dK_x \cong n^2[s.e.(\hat{\Lambda}_\tau)]^2,$$

$$\hat{\theta}^2 C_\tau \cong \hat{\theta}^2 (R_\tau - 2L_\tau),$$

$$2\sum_{h=1}^{n} \int_0^\tau \frac{(n-L_x)\,dK_h(x)}{s_h\left\{s - \sum_h s_h K_h(x)\right\}} \cong \frac{2}{m_a} B_\tau \cong \frac{2n}{m_a}\hat{\Lambda}_\tau.$$

The approximations given for each of the first three estimates are exact when at

least one household escapes infection and no households are completely infected. However, they will be good approximations as long as the disease is not highly infectious, that is as long as L_τ is small compared with n. The same is true for the fourth estimate when all households are of the same size. The approximations for the fourth estimate rely on the additional approximation of setting $s_h = m_a$, for affected households, when the households are of varying sizes.

With these estimates we may compute

$$\text{s.e.}(\hat{\theta}) = \left\{\hat{V}[M_1^*(\tau) - M_1^{\wp}(\tau) - M_2^*(\tau)]\right\}^{1/2} / C_\tau.$$

Suppose the disease is such that the infectious period has the same (constant) duration for each infective. In this case $\pi = 1\text{-}e^{-\theta}$ is the probability that a given susceptible makes at least one infectious contact with a given infective from the same household during the latter's infectious period. Haber et al. (1988) refer to π as the secondary attack rate. In the case of a constant infectious period one estimates this by

$$\hat{\pi} = 1 - e^{-\hat{\theta}} \text{ and s.e.}(\hat{\pi}) = e^{-\hat{\theta}} \text{ s.e.}(\hat{\theta}).$$

An important difference is that, with a constant infectious period, one needs to set $V[M_2^*(\tau)] = 0$ in the computation of s.e. $(\hat{\theta})$.

4. Application to Influenza Data. Data on the sizes of outbreaks in Tecumseh households over the 1977-78 influenza epidemic season are given in Table 1. These are taken from Table 1 of Haber et al. (1988).

4.1. Community Acquired Infection. There are $K_\tau = 76$ affected households. The average size of affected households is $m_a = 206/K_\tau = 2.7105$. Substituting these values and $n^* = 616/m_a = 227.26$ into (2.5) and (2.6) gives

$$\hat{\Lambda}_\tau = 0.150 \text{ and s.e.}(\hat{\Lambda}_\tau) = 0.017.$$

This leads to

$$\hat{P}_\tau = 1 - e^{-\hat{\Lambda}_\tau} = 0.139 \text{ and s.e.}(\hat{P}_\tau) = e^{-\hat{\Lambda}_\tau} \text{ s.e.}(\hat{\Lambda}_\tau) = 0.015,$$

which compares exceptionally well with the maximum likelihood estimation 0.140 ± 0.015 given by Haber et al. (1988, Table 1).

It is instructive to compute an estimate of Λ_τ based only on the data from households of size j, for each j=1,2,3 and 4. The four estimates of Λ_τ obtained in this way are shown in column 4 of Table 2, with the corresponding standard errors in column 5. These estimates should be inspected for large differences and trends with increasing household size. Such anomalies would cast doubt on the underlying assumptions. No patterns or large differences are apparent in the current data set (unless there is a good reason why households of size two seem less prone to community acquired infection). The weighted average of the estimates of Λ_τ given in Table 2, with weights inversely proportional to the estimated variances, is 0.124 ± 0.014.

4.2. The Within-Household Infection Potential. We now concern ourselves only with households of size two, three and four. For these households we have n=200, s=538, K_τ =63, L_τ=9 and m_a=193/K_τ. These lead to

$$A_\tau = 41.5,$$

$$B_\tau \cong \frac{n}{m_a} ln\left(\frac{n^* + 1/2}{n^* - K_\tau + 1/2}\right) = 28.9034$$

and $C_\tau \cong R_\tau - 2L_\tau = 79,$

from which we find

$$\hat{\theta} = (A_\tau - B_\tau)/C_\tau = 0.159.$$

The corresponding standard error is computed to be

$$\text{s.e.}(\hat{\theta}) = [21.875 + 13.4845 + 2.0085 - 18.8696]^{1/2}/79 = 0.054 .$$

From these we obtain

$$\hat{\pi} = 1 - e^{-\hat{\theta}} = 0.147 \text{ and s.e.}(\hat{\pi}) = 0.044 .$$

These compare reasonably well with the maximum likelihood estimate 0.155 ± 0.035 given by Haber et al. (1988, Table 1). Note that in the calculation of s.e.($\hat{\pi}$) we deleted the number 2.0085 from the above numerical expression for s.e.($\hat{\theta}$). This is appropriate when it is assumed that there is no variation in the duration of the infectious period.

When the within-household infection potential is estimated separately for households of size 2, 3 and 4 one obtains the estimates shown in Table 3. These should be inspected for large differences and trends, which provides a rough

check of model assumptions. Although the estimates of θ decrease as m increases it is unwise to read too much into this because the standard errors are relatively large . The weighted average of these estimates of θ is also given in Table 3. This is seen to be close to the pooled estimate given above.

Acknowledgement. The author gratefully acknowledges support from the Danish Research Academy and the Statistical Research Unit, University of Copenhagen.

References

[1] Andersen, P.K. & Borgan, ϕ. (1985). Counting process models for life history data: a review. *Scand. J. Statist.* **12**, 97-158.

[2] Becker, N.G. (1979). An estimation procedure for household disease data. *Biometrika* **66**, 271-277.

[3] Becker, N.G.(1981). The infectiousness of a disease within households. *Biometrika* **68**, 133-141.

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

[5] Haber, M., Longini, I.M. & Cotsonis, G.A. (1988). Models for the statistical analysis of infectious disease data. *Biometrics* **44**, 163-173.

[6] Rao, C.R. (1973). Linear Statistical Inference and it Applications (2nd edition). New York: Wiley.

Table 1. Observed frequencies of outbreak sizes in Tecumseh during the 1977-1978 influenza epidemic season.

	Initial number of susceptibles per household			
Number Infected	1	2	3	4
0	65	88	27	22
1	13	14	15	9
2		4	4	9
3			4	3
4				1
Total	78	106	50	44

Table 2. Requisite data and estimates for community acquired infection.

Household size	Frequency	Affected households		
		K_τ	$\hat{\Lambda}_\tau$	s.e. $(\hat{\Lambda}_\tau)$
1	78	13	0.181	0.050
2	106	18	0.093	0.022
3	50	23	0.203	0.043
4	44	22	0.170	0.037
Total	278	76		

Table 3. Estimates, and associated standard errors, of the within-household infection potential for households of size 2, 3 and 4.

Household size, m	Infection potential $\hat{\theta}$	Standard error s.e. $(\hat{\theta})$
2	0.228	0.156
3	0.205	0.102
4	0.140	0.056
Weighted average	0.162	0.047