Tests to Detect Clustering of Infected Individuals within Families
Author(s): Tom Britton
Source: *Biometrics*, Vol. 53, No. 1 (Mar., 1997), pp. 98-109
Published by: International Biometric Society
Stable URL: http://www.jstor.org/stable/2533100
Accessed: 18/08/2009 03:45

BIOMETRICS 53, 98–109
March 1997

Tests to Detect Clustering of Infected Individuals within Families

Tom Britton

Mathematical Statistics, Department of Mathematics,
Stockholm University, S-106 91 Stockholm, Sweden

SUMMARY

A new simple test to detect within-family clustering of infected individuals is proposed. The test is derived as the score test for several different parametric models designed to allow an increased within-family infectivity. The new test is compared with other tests proposed to detect the same type of clustering caused by increased within-family infectivity. Applications of household disease data and simulations are used to illustrate the theory.

1. Introduction

In epidemiology the transmission dynamics of specific infectious diseases is an important research area. One particular concept in this field is the secondary attack rate (SAR). SAR quantifies the extra probability of getting infected due to other infected individuals of the same "unit." In applications a unit might, for example, be a family, a household, or a village; we shall use the term family. A high SAR would typically result in clustering of infected individuals within families.

There is a vast amount of literature on this subject. For example, Longini and Koopman (1982) treat a parameterized model (the LK model) and describe how to obtain parameter estimates from data. In Addy, Longini, and Haber (1991) the LK model is extended to a multitype population. Fraser (1983) and Walter (1974) propose tests for the hypothesis of no extra within-family infectivity, i.e., that SAR = 0. The tests are motivated by their simple structure and their intuitive interpretation. Schork (1994) proposes a different test of the same hypothesis based on the parameter estimates of the LK model.

In the present paper we suggest a new simple test of the hypothesis that SAR = 0. The test is similar to the tests of Fraser (1983) and Walter (1974). The test is obtained as the asymptotic score test of the hypothesis of no extra within-family infectivity in a simple epidemic model. This means that, for a large population and small deviation from the null hypothesis, the suggested test is the most powerful test. In the Appendix we outline the proof giving the suggested test statistic. An extension of the epidemic model treated in the present paper is analyzed in a recent paper by Ball, Mollison, and Scalia-Tomba (1996). It turns out that the same test is the asymptotic score test for this extended model. If the LK model is modified in a natural way, the corresponding local test is also identical to the one presented here. The different epidemic models are described in Section 2 and the corresponding tests are presented in Section 3.

If the disease is known to spread through "close" person-to-person contact, the hypothesis would of course most likely be rejected and the test would be uninformative. On the other hand, for several infectious diseases, the mechanism by which the diseases spread is unknown. Other common ways of spread are on a more "global" level, such as through a water source shared by the whole population or by air. In the model we will study, the disease spreads only through person-to-person contact on a global level under the null hypothesis—a global "external source" would need a different model. However, in the test procedure, we will condition on the total number of infected individuals and, given this number, each configuration has equal probability under the null hypothesis. We would have the same conditional distribution if the disease were spread through a global external source

Key words: Clustering within families; Epidemic model; Score test.

such as a common water supply or by air. Thus, the test presented in this paper can be interpreted as a test to find out whether the disease spreads through close person-to-person contact, as opposed to any spreading mechanism on a global level.

The statistical data needed for the test are, besides family sizes, the final numbers of infected individuals in each family. The times of infection (onset times) are assumed unobserved and hence not used. Examples of such data are the influenza data collected in Tecumseh and Seattle (Longini et al., 1982; Haber, Longini, and Cotsonis, 1988; Addy et al., 1991; Schork, 1994), the Asian influenza data (Longini and Koopman, 1982), the spread of *Trypanosoma cruzi* among 40 Brazilian households (Smith and Pike, 1976; Fraser, 1983), and a study of Lassa Fever in Sierra Leone (Fraser, 1983).

In Section 4 we compare the tests of Section 3 on some examples of data. We also present some simulations to compare power properties.

2. The Standard Epidemic Model for a Population of Families

2.1 *Model Specification*

The model we will study, which we denote the Standard Epidemic Model for a Population of Families, is an extension of what is known as the General Epidemic Model. As the epidemic evolves, individuals may pass through three stages: susceptible (meaning that the individual is susceptible to infection), infective (if an individual is infected, he/she starts spreading the disease to other individuals), and removed (after some time the infective individual stops spreading the disease further and becomes immune). Models of this kind are called SIR-epidemic models (see Lefèvre, 1990, for a nice exposition of such models).

The dynamics are as follows. An individual who gets infected remains infective for an exponentially distributed time. Without loss of generality, we assume the mean length is 1 time unit. While infective, the individual makes contact at rate λ with someone chosen at random from the remaining $N-1$ individuals (N denotes the population size). Independent of this contact process, he/she makes contact at rate δ with someone chosen randomly from his/her own family. Whenever a contact is made with a susceptible individual, this individual becomes infected and infective; otherwise, nothing happens. All contact processes, random selection numbers, and infectious periods are defined as mutually independent.

In the Appendix, and more rigorously in Britton (1996), this model is analyzed using theory for counting processes, where intensities play an important role. The above description implies the following intensities (rates) with which individuals change state.

1. Susceptible→infective. Consider a given susceptible individual in a family of size f in a population of size N. If there presently are I infective individuals in the whole population and i infective individuals in the same family as our susceptible individual, then this individual becomes infected with intensity $\lambda I/(N-1) + \delta i/(f-1)$.
2. Infective→removed. An infective individual becomes removed with intensity 1, since the infectious period is exponentially distributed with mean 1.

The epidemic stops the first time there are no infective individuals in the population because then no one can get infected. When $\delta = 0$, there is no increased within-family infectivity and the model reduces to the General Epidemic Model.

2.2 *Relations to Other Models*

Ball et al. (1996) treat an extension of the model described in the previous subsection, allowing a general distribution for τ, the length of the infectious period, whereas here this time period is assumed to follow the exponential distribution. The same asymptotic situation is also treated, so results from Britton (1996) that will be used can just as well be derived from Ball et al. (1996).

The LK model (Longini and Koopman, 1982) is a different two-parameter model. The parameters are B and Q, where B is the probability of a susceptible individual escaping infection from outside the family during the course of the epidemic and Q is the probability of escaping infection from an infected individual in the same family during the time this individual is infective.

In the LK model the probability of getting infected by someone within the family depends on how many family members get infected. On the other hand, the probability of getting infected from outside the family, B, is independent of how many individuals outside the family become infected. This assumption simplifies calculations; in particular, all families behave independently so the log likelihood is a sum of independent terms. On the other hand, it might not be a realistic assumption for particular infectious diseases. Another assumption in the LK model is that the two events of an infected individual infecting two different family members are independent events. If the length the infectious period for a particular disease is considered random, these two events will most likely be positively correlated. In the model of this paper, neither of these assumptions is true.

The LK model can be modified so that Q depends on the family size f, that is $Q = Q_f$. For example, in a family of size two, the probability of infecting your family mate might be larger than if the family was larger, say of size six. In Section 3.3 we will treat the choice $Q_f = 1 - \delta/(f-1)$, where δ should be interpreted as the average "infectivity mass" which is spread uniformly to the $f-1$ family companions. We will see that this modification of the LK model resembles more closely the present model, in particular it gives rise to the same score test of the hypothesis $\delta = 0$.

Addy et al. (1991) generalize the LK model so that individuals may be different types and, by letting the length of the infectious period τ follow an arbitrary distribution, may possibly be type-specific. Instead of Q, the intensity parameters $\{\beta_{i,k}\}$ are introduced, where $\beta_{i,k}$ is the intensity with which a susceptible i-type individual is infected by an infective k-type individual in the same family. The different types might, for example, depend on age and/or sex. The parameter B has the same interpretation in Addy et al. (1991) as in the LK model.

3. Testing for Clustering

3.1 *The Standard Epidemic Model for a Population of Families*

For the model described in Section 2.1, $\delta = 0$ clearly corresponds to the case SAR $= 0$, which implies that there is no tendency for clustering of infected individuals within families. This is the hypothesis for which we will construct a test.

In a fairly large population, the exact distribution of the final outcome of infected individuals (initiated by a small proportion of infectives) is complicated, so a test based on the likelihood is hard to derive. Using theory for Markov counting processes (cf. Ethier and Kurtz, 1986), asymptotic properties of the model have been derived (Britton, 1996). In the Appendix we outline how this is done and give an heuristic argument for the score statistic presented in (3.1).

Let $N_{f,i}$ denote the number of families of size f that had exactly i infected individuals at the end of the epidemic, and denote the largest family size by f_{max}. For a population with a large number of families, it is shown that the distribution of the vector with elements $\{N_{f,i}; 0 \leq i \leq f, 1 \leq f \leq f_{max}\}$ is approximately (i.e., asymptotically) multivariate normal with mean vector and covariance matrix depending on λ and δ.

Using the corresponding normal densities, this allows us to approximate the likelihood ratio, which we denote $L(\delta)/L(0)$, suppressing the dependence on λ. For general δ, the mean vector and covariance matrix are only given implicitly and, beside this, the likelihood ratio is known to depend on which particular δ is considered. This implies that no uniformly most powerful test of the hypothesis $\delta = 0$ exists. We proceed by constructing the score test, the test that maximizes the power for small deviations from the null hypothesis (cf. Cox and Hinkley, 1974, pp. 113). The motivation for this approach is that, for a large deviation from the null hypothesis, the hypothesis will most likely be rejected anyway. The score test tells us to reject the hypothesis whenever $l'(0) = L'(0)/L(0)$ is large ($l(\delta)$ is the log likelihood). The distribution of $l'(0)$ depends on the nuisance parameter λ. We therefore condition on the total number of infected individuals, $R = \Sigma_{f=1}^{f_{max}} \Sigma_{i=0}^{f} iN_{f,i}$, this being a sufficient statistic for λ under the null hypothesis. Thus, we may omit terms in $l'(0)$ only depending on data through R, as well as terms independent of data. Finally, we replace the unknown parameter $p = p(\lambda)$, which is the asymptotic probability of being infected, by the observed proportion infected, $\tilde{p} = R/N$. In the Appendix, where an outline of this analysis is given, the resulting test statistic is seen to be

$$T = \tilde{p}N_{1,1} + \sum_{f=2}^{f_{max}} \sum_{i=0}^{f} \frac{i(i-1)}{f-1} N_{f,i}. \tag{3.1}$$

From the normal approximation mentioned above it can be shown that, conditioned on $\tilde{p}$, T is approximately normally distributed with mean $\mu = \tilde{p}^2 N$ and variance $\sigma^2 = 2\tilde{p}^2\tilde{q}^2\Sigma_{f=2}^{f_{max}} [f/(f-1)]n_f + \tilde{p}^3\tilde{q}(n_1 - n_1^2/N)$, where n_f is the number of families of size f, $N = \Sigma_f fn_f$ is the total population size, and $\tilde{q} = 1 - \tilde{p}$.

We are approximating the distribution of a discrete statistic with the normal distribution. As with the binomial distribution, this approximation can be improved with a continuity correction.

A natural choice is to let the continuity correction be half the size of the smallest difference of two possible T-values. If the different family sizes are $f_1, \dots, f_r$, it is easy to show that this gives the continuity correction $c = 1/\text{lcm}\{f_1 - 1, \dots, f_r - 1\}$ if there are no single-individual families and $c = 1/\text{lcm}\{f_1 - 1, \dots, f_r - 1, 2k\}$ otherwise, where lcm denotes least common multiple and k is defined from $\tilde{p}$: $\tilde{p} = j/k$ such that j and k are relatively prime. For example, if the different family sizes are 3, 4, and 5 we subtract one from each size and write this number as a product of primes: $3 - 1 = 2^1$, $4 - 1 = 3^1$, $5 - 1 = 2^2$, and we conclude that $\text{lcm}\{2, 3, 4\} = 2^2 \cdot 3^1 = 12$.

Test procedure. Reject the null hypothesis of no extra within-family infectivity (i.e., SAR $= 0$) if $(T - c - \mu)/\sigma$ is significantly large compared to the standard normal distribution.

The first term of T in (3.1) might seem peculiar: Does a large number of infected individuals in single-individual families indicate that there is evident within-family infectivity? Also, other proposed tests for clustering (see Section 3.4) are of the same type as T but leave out the first term. However, its presence can be motivated; which is most easily done with a simple example.

Consider two different epidemics in identical, very small populations: one single-individual family, one size-three family, and one family of size four. Each of the epidemics result in four infected individuals, making μ, σ, and c identical for the two epidemics. In the first epidemic, the single individual and three individuals in the family of size four were infected, making $T = 2.5$. In the second epidemic, one individual in the size-three family and three individuals in the family of size four were infected, so $T = 2$ for this epidemic. Is the second epidemic less likely to infect within families at a higher rate, as indicated by the T-values? Yes, because in the second epidemic the one infected individual in the family of size three failed to infect his/her family mates, whereas in the first epidemic the single individual had no family member to infect. The test statistic leaving out the first term (called T_1 in Section 3.4) would not distinguish these cases.

As mentioned in the Introduction, the model in Ball et al. (1996) is an extension of the model of this paper. It may be adjusted so that, during the infectious period, an infective attempts to infect someone in the family at rate δ irrespective of the family size. In a family of size f a given susceptible is thus infected by a given infective at rate $\delta/(f - 1)$. Just like the case for the model of the present paper—and contrary to the LK model and the model of Addy et al. (1991)—there is no exact closed form for the distribution of the final outcome for this model due to the dependence between families. However, in Ball et al. (1996) the asymptotic distribution of the final outcome of the epidemic is derived using the model of Addy et al. (1991) with a homogenous population but arbitrary distribution of τ, the length of the infectious period. It is shown that, if the parameter B (the probability of escaping infection from outside the family) in the model of Addy et al. (1991) is replaced by the probability of escaping infection from the expected infectious force generated by the whole population outside one's family, then this gives the asymptotic distribution of the final outcome in the model of Ball et al. (1996). If p denotes the asymptotic probability of ever getting infected, $\text{E}(\tau)$ is the mean length of an infectious period, and λ is the rate an individual tries to infect individuals outside the family, then the expected infectious force is $\lambda p \text{E}(\tau)$ and the probability of escaping this force is $e^{-\lambda p \text{E}(\tau)}$, so this is what should replace the parameter B.

Ball et al. (1996) also derive a recursive method to obtain $q_{f,i}(\delta)$ (the asymptotic probability for a family of size f of having i infected individuals) using Gontcharoff polynomials. From this it can be verified that the score statistic of their model is equivalent to (3.1). Just as with the present model, the results are based on approximations relying on a large number of families.

3.2 *The LK Model*

The probability that exactly i individuals will get infected in a family of size f, denoted by $p_{f,i}$, is derived recursively as follows. For a fixed subgroup of size i in the family, the probability that all individuals in the subgroup and no one else in the family will get infected is $p_{i,i}(BQ^i)^{f-i}$, because the subgroup has to be completely infected without the "help" of the other family members, and the remaining individuals all have to escape infection from outside the family and from the i infected individuals. Because there are $\binom{f}{i}$ distinct subgroups of size i, we have $p_{f,i} = \binom{f}{i} p_{i,i} B^{f-i} Q^{i(f-i)}$.

If we substitute $1 - \delta$ for Q it can be shown that for small δ we have

$$p_{f,i} = \binom{f}{i}(1 - B)^i B^{f-i} + \delta \binom{f}{i}(1 - B)^i B^{f-i} \left(\frac{i(i-1)}{(1-B)} - i(f-1) \right) + o(\delta). \qquad (3.2)$$

The parameter δ is different than before, but $\delta = 0$ still corresponds to no within-family infectivity (SAR $= 0$).

For the same reason as in Section 3.1, we test the hypothesis using the score statistic $(\partial/\partial\delta)\, l(B, \delta)|_{\delta=0}$. As mentioned in Section 2.2, the LK model implies that families behave independently, so the log likelihood splits into a sum. From this and using (3.2), it is easy to verify that

$$\frac{\partial}{\partial\delta}l(B,\delta)\Big|_{\delta=0} = \sum_{f=1}^{f_{max}}\sum_{i=0}^{f}\left(\frac{i(i-1)}{(1-B)} - i(f-1)\right)N_{f,i}.$$

The expression above contains the unknown nuisance parameter B and, just like with λ in the previous subsection, the total number of infected individuals $\Sigma_{f=1}^{f_{max}}\Sigma_{i=0}^{f} iN_{f,i}$ is sufficient for B (when $\delta = 0$). Thus, we shall condition on the observed value of this statistic. A test based on the sum above is then equivalent to $\Sigma_{f=1}^{f_{max}}\Sigma_{i=0}^{f} i(i - f(1 - B))N_{f,i}$. Finally, we replace the unknown parameter $1 - B$, the probability of getting infected when H_0 is true, by the observed proportion infected $\tilde{p}$ to end up with the test statistic

$$T_{LK} = \sum_{f=1}^{f_{max}}\sum_{i=0}^{f} i(i - f\tilde{p})N_{f,i}. \tag{3.3}$$

This statistic is also approximately normally distributed. Under the null hypothesis, the conditional mean of T_{LK} is $\mu = \tilde{p}\tilde{q}N$. The conditional variance is $\sigma^2 = \tilde{p}^2\tilde{q}(1+\tilde{q})N_{(2)} + \tilde{p}^3\tilde{q}(N_{(3)} - N_{(2)}^2/N)$, where $N_{(i)} = \Sigma_f f(f-1)\cdots(f-i+1)n_f$.

This test can, in some cases, show a surprising feature. For example, suppose there are only two families, one of size two and the other of size six, that together have four infected individuals. This fixes the moments above. If the smaller family had one infected individual and the larger family had three infected individuals, we would have $T_{LK} = 0$. On the other hand, if both individuals in the smaller family were infected and two individuals in the larger family were infected, $T_{LK} = 0$ as well. However, in the former case both families have 50% infected individuals, whereas in the latter case the small family has 100% infected individuals and the large family has only 33% infected individuals, indicating a clustering in the second case but not in the first. This example could be an argument for why the probability of infecting a given individual of a family should be larger in a small family than in a large family in the model, as is the case for the model of this paper and the model of the next subsection, but not for the LK model. Of course, what is preferred depends on the particular disease of interest. Fraser (1983) uses a similar example to motivate his test, as opposed to the test in Walter (1974); see Section 3.4.

The test using T_{LK} was derived as the score test for the hypothesis that $\delta = 0$. Another test is, of course, the maximum likelihood ratio (MLR) test, $2(l(\hat{B},\hat{\delta}) - l(\hat{B}_0, 0))$, where l is the log likelihood, $\hat{B}$ and $\hat{\delta}$ are the ML estimates, and $\hat{B}_0$ is the ML estimate under the restriction that $\delta = 0$. Besides a constant, the log likelihood is $\Sigma_{f=1}^{f_{max}}\Sigma_{i=0}^{f}\log(p_{f,i})N_{f,i}$, and $p_{f,i}$ is given by (3.2). Numerically, it is possible to find the value of the test statistic after some programming and computing. Under the null hypothesis, MLR is approximately distributed as a (1/2,1/2)-mixture of χ^2 with one degree of freedom and a point mass at 0; the point mass at 0 comes from the fact that the null hypothesis ($\delta = 0$) is on the boundary of the parameter space. This is the test procedure Schork (1994) proposes. In Section 4 we shall use this test (among others) on some examples.

3.3 *The Modified LK Model*

Now we modify and reparameterize the LK model as described in Section 2.2 to let Q, the probability of escaping infection from a given infected of the same family, depend on the family size: $Q = Q_f = 1 - \delta/(f-1)$ for $f \geq 2$, where δ is our new parameter. We get a result corresponding to (3.2) using recursive methods similar to those leading to (3.2). However, now the probability that all i individuals are infected in a size-i subgroup of a family depends on the family size and is hence not equal to $p_{i,i}$, as was the case in the original LK model. In the modified model, an infective in a family with f individuals infects a given susceptible in the same family at rate $\delta/(f-1)$. If we write $p_{f,i}(\delta)$ to emphasize what δ is used, it should be clear that $p_{f,i}(\delta) = \binom{f}{i}p_{i,i}[\delta(i-1)/(f-10)](B[1-\delta/(f-1)]^i)^{f-i}$ because when δ is multiplied by $(i-1)/(f-1)$ the infectivity rate in families with i individuals is equal to the infectivity rate in families of size f when δ is the parameter. It can be shown that this implies

$$p_{f,i}(\delta) = \binom{f}{i}(1-B)^iB^{f-i} + \delta\binom{f}{i}(1-B)^iB^{f-i}\left(\frac{i(i-1)}{(1-B)(f-1)} - i\right) + o(\delta)$$

for $f \geq 2$. For $f = 1$ we have $p_{1,0} = B = 1 - p_{1,1}$ (there is no one to infect in a single-individual family). We base our test on the score, which for the modified LK model equals $\Sigma_{f=2}^{f_{max}} \Sigma_{i=0}^{f} (i(i-1)/(1-B)(f-1) - i)N_{f,i}$.

As earlier, $\Sigma_{f=1}^{f_{max}} \Sigma_{i=0}^{f} iN_{f,i}$ is sufficient for B (when $\delta = 0$). We therefore condition on the observed value of this statistic, which, subtracted by $N_{1,1}$, equals the negative term above. So if we add $\Sigma_{f=1}^{f_{max}} \Sigma_{i=0}^{f} iN_{f,i}$, multiply with $1 - B$, and replace $1 - B$ by $\tilde{p}$, we obtain the test statistic

$$T = \tilde{p}N_{1,1} + \sum_{f=2}^{f_{max}} \sum_{i=0}^{f} \frac{i(i-1)}{f-1} N_{f,i},$$

i.e., the same test as in Section 3.1.

As with the original LK model, another test is the maximum likelihood ratio (MLR) test, $2(l(\hat{B}, \hat{\delta}) - l(\hat{B}_0, 0))$, which under the null hypothesis is approximately distributed as a (1/2,1/2)-mixture of χ^2 with one degree of freedom and a point mass at 0.

3.4 *Other Tests*

Walter (1974) and Fraser (1983) propose tests that are similar to (3.1). They differ from (3.1) by leaving out the first term and having a different denominator in the sum. The test statistic in Walter (1974) is

$$W = \sum_{f=2}^{f_{max}} \sum_{i=0}^{f} i(i-1)N_{f,i}. \tag{3.4}$$

The mean and variance to be used when calculating the level of significance is $\mu = \tilde{p}^2 N_{(2)}$ and $\sigma^2 = 2\tilde{p}^2\tilde{q}^2 N_{(2)} + 4\tilde{p}^3\tilde{q}(N_{(2)} + N_{(3)} - N_{(2)}^2/N)$, using notation defined after (3.3). A natural continuity correction is 1, since W is always an even integer.

Fraser (1983) introduces another test of the same type as (3.4). He motivates why (3.4) should be modified with the arguments used against T_{LK} in Section 3.2. The test statistic is

$$F = \sum_{f=2}^{f_{max}} \sum_{i=0}^{f} \frac{i(i-1)}{f} N_{f,i}. \tag{3.5}$$

Under the null hypothesis, the mean of F is $\mu = \tilde{p}^2(N - n)$ and the variance is $\sigma^2 = 2\tilde{p}^2\tilde{q}^2(n - \Sigma_f (n_f/f)) + 4\tilde{p}^3\tilde{q}(\Sigma_f (n_f/f) - n^2/N)$, where $n = \Sigma_f n_f$ is the number of families. A natural continuity correction for F is $1/\text{lcm}\{f_1, \ldots, f_r\}$, by the same arguments as for T.

In Section 3.1 we discussed the possibility of leaving out the first term in (3.1), giving a test similar to the ones above. This new test is

$$T_1 = \sum_{f=2}^{f_{max}} \sum_{i=0}^{f} \frac{i(i-1)}{f-1} N_{f,i}. \tag{3.6}$$

Under the null hypothesis, the mean of T_1 is $\tilde{p}^2(N - n_1)$ and the variance is $2\tilde{p}^2\tilde{q}^2 \Sigma_{f=2}^{f_{max}} [f/(f-1)]n_f$. The natural continuity correction for T_1 is $1/\text{lcm}\{f_1 - 1, \ldots, f_r - 1\}$.

4. Examples of Infectious Disease Data

Now we shall use the tests of Section 3 on three sets of data taken from Longini and Koopman (1982), Fraser (1983), and Longini et al. (1982). The tests that will be used are the ones based on T, T_{LK}, W, F, and T_1, given by (3.1), (3.3), (3.4), (3.5), and (3.6), respectively, and the MLR tests for the original LK model, as well as for the modified version.

Example 1. This example concerns the outcome of the Asian influenza data of Sugiyama among 42 families of size 3 (see Longini and Koopman, 1982). Among the 42 families, 29 had no infected individuals, 9 families had 1 infected individual, 2 families had 2 infected individuals, and, finally, 2 families had all 3 individuals infected.

When all families are of the same size, it is easily verified that the tests based on T, W, F, and T_1 are equivalent. The same is true for the two MLR tests, so there are only three different tests, the tests based on T and T_{LK} and MLR-LK, say. For this data we have $T = 8$, $\mu = 2.865$, the continuity correction $c = 1/2$, and $\sigma^2 = 2.07$, giving the p-value 0.00063. If we use (3.3) on the Asian influenza data we have $T_{LK} = 26.40$, $\mu = 16.135$, and $\sigma^2 = 8.265$, implying that $p = 0.00018$.

By maximizing the likelihood numerically under the general hypothesis as well as under the null hypothesis, it is found that the MLR statistic equals 7.82 for the original LK model. Compared with a (1/2,1/2)-mixture of χ^2 with one degree of freedom and a point mass at 0, this gives the p-value 0.00264.

Example 2. In Table 1 we find the made-up data from Fraser (1983). There are 28 individuals in 14 families. All six tests are different and the resulting p-values are shown in Table 3, where it is seen that they vary between 0.029 and 0.117.

Table 1
Example 2. *Fraser* (1983). *Observed number of families.*

	Number of infected			
Family size	0	1	2	3
1	2	2	—	—
2	3	1	2	—
3	1	1	1	1

Example 3. Our last example shows data concerning influenza B collected in Seattle (Longini et al., 1982) (see Table 2). Fifty-six out of 259 individuals were reported infected; family sizes varied between 1 and 5. In Table 3 we see that p-values range from $6 \cdot 10^{-5}$ for T_{LK} to $3 \cdot 10^{-4}$ for the MLR statistic from the modified LK model.

In Table 3 we summarize the results from the three examples. It is seen that the different tests give different results, which is quite natural. The difference in magnitude seems to be a factor between 5 and 10. For the data in example 2 (Table 1), some tests would have rejected the null hypothesis at the 5% level, whereas others would not.

Table 2
Example 3. *Longini et al.* (1982). *Influenza B,* 1975–1976, *in Seattle. Observed number of families.*

	Number of infected					
Family size	0	1	2	3	4	5
1	9	1	—	—	—	—
2	12	6	2	—	—	—
3	18	6	3	1	—	—
4	9	4	4	3	0	—
5	4	3	0	2	0	0

Table 3
Observed p-values for different tests based on data from examples 1–3

	Data		
Test	Example 1	Example 2	Example 3
T	0.0006	0.0312	0.00011
MLR-mod LK	0.0026	0.0292	0.00029
T_{LK}	0.0002	0.0357	0.00006
MLR-LK	0.0026	0.0395	0.00024
W	0.0006	0.1168	0.00022
F	0.0006	0.0681	0.00006
T_1	0.0006	0.0595	0.00008

To see which test has the best performance for a specific set of parameter values in a given finite population when the Standard Epidemic Model for a Population of Families is the true model, we have performed some simulations. The population was chosen to consist of 40 families, 10 families each with size 1, 2, 3, and 4; hence, the total population was 100 individuals. The epidemic was simulated under two different sets of parameter values, one when the null hypothesis was true ($\delta = 0$) and the other when it was not. Simulations with less than 10% (= 10 individuals) infected were disregarded because approximations rely on a positive fraction being infected. In both cases, 1000 epidemics with at least 10% infected were simulated.

The specific choices of parameter values used with the null hypothesis were $\lambda = 1.5$ and $\delta = 0$. The other parameter set was $\lambda = 1.05$ and $\delta = 0.45$. These choices are, of course, very arbitrary. They were chosen as follows. In both cases $\lambda + \delta = 1.5$. This was chosen so that, on the average, about 50% would be infected. The relation $\lambda = 1.05$ and $\delta = 0.45$ was chosen after some preliminary simulations to make sure that the tests would reject the null hypothesis approximately 50% of the time.

In Table 4 we have listed the proportion of H_0-simulations that would reject the null hypothesis at some common significance levels. For a test not to be skew, these proportions should approximately equal the corresponding significance level.

In Table 5 the corresponding proportions are listed, now for the simulations with $\lambda = 1.05$ and $\delta = 0.45$, i.e., the null hypothesis is not true. The higher the proportions the more powerful is the test. We see that T_1 is most powerful among the studied tests for the 1000 simulations performed. However, in Table 4 it is seen that T_1 is skew since the observed proportion with small p-values is larger than expected under the null hypothesis. The reason for this is most likely that the normal approximation does not work perfectly when the population consists of only 100 individuals. (The MLR test for the LK model and W seem to be somewhat skew to the other direction.) Among the remaining tests, we see in Table 5 that T is most powerful, and from Table 4 it seems that the level of significance agrees quite well with the observed frequencies. T is more powerful than the corresponding MLR-test, although the difference is not severe. Of course no general conclusions can be drawn from these simulations, but they speak in favor of the new test T.

Table 4

One thousand simulations under H_0. Proportion of simulations rejecting H_0 for some common significance levels.

	Level of significance			
Test	0.1	0.05	0.01	0.005
T	0.097	0.051	0.013	0.006
MLR-mod LK	0.094	0.045	0.009	0.004
T_{LK}	0.096	0.053	0.011	0.009
MLR-LK	0.084	0.041	0.006	0.003
W	0.071	0.041	0.010	0.005
F	0.092	0.051	0.013	0.008
T_1	0.118	0.077	0.029	0.021

Table 5

One thousand simulations under H_A. Proportion of simulations rejecting H_A for some common significance levels.

	Level of significance			
Test	0.1	0.05	0.01	0.005
T	0.737	0.641	0.426	0.347
MLR-mod LK	0.730	0.617	0.370	0.285
T_{LK}	0.722	0.612	0.389	0.311
MLR-LK	0.713	0.573	0.329	0.247
W	0.617	0.468	0.278	0.247
F	0.714	0.603	0.391	0.313
T_1	0.767	0.681	0.496	0.426

5. Discussion

For some infectious diseases, individuals may not be equally likely to get infected. For example age and/or sex might be an important explanatory factor for the disease. For the models treated in this paper, this type of heterogeneity is not permitted, which is clearly a drawback. The different models can be generalized to so-called multitype populations, allowing different types of individuals. A test statistic and its approximate distribution could be obtained using arguments similar to those of the present paper. Unfortunately, the structure of the statistic and its moments are no longer simple and hence are not easily used to test for clustering within families. By numerically maximizing the log likelihood in the model of Addy et al. (1991), an MLR test can be derived for the case of a multitype population.

Commenges et al. (1994) derive the score test of the hypothesis of homogeneity between groups (or equivalently families) when the alternative hypothesis is that random effects, in a logistic regression model, of individuals belonging to the same group are positively correlated. The resulting score statistic for their model is similar to the ones of Section 3. In particular, if all explanatory variables are identical, their score statistic is $\Sigma_{f=1}^{f_{max}} \Sigma_{i=0}^{f} i(i - 2f\tilde{p})N_{f,i}$, which is much like T_{LK} defined in (3.3). Although similar, the technique used to derive their statistic is different because the models are fundamentally different. In their model, all individuals behave independently conditional on the random effects, whereas in the present model individuals behave in a genuinely correlated fashion.

Acknowledgements

This paper is part of the author's Ph.D. thesis, written under guidance of Åke Svensson, to whom the author is grateful. Thanks also to two referees and an associate editor for comments that helped improve the paper considerably. This work was supported in part by The Bank of Sweden Tercentenary Foundation.

Résumé

Un nouveau test simple pour détecter des classes intra-fratrie d'individus infectés est proposé. Le test est développé comme un score-test pour différents modèles paramétriques mis au point pour tenir compte d'un accroissement du risque d'infection intra-fratrie. Le nouveau test est comparé à d'autres tests proposés pour détecter le même type classification induit par un tel accroissement du risque d'infection. Des applications à des données de maladie de fratrie ainsi que des simulations sont utilisées pour illustrer la théorie.

References

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

Ball, F. G., Mollison, D., and Scalia-Tomba, G. (1996). Epidemics with two levels of mixing. *Annals of Applied Probability,* in press.

Britton, T. (1996). Epidemics with heterogeneous mixing: Stochastic models and statistical tests. Ph.D. thesis, Stockholm University, Stockholm.

Commenges, D., Letenneur, L., Jacqmin, H., Moreau, T., and Dartigues, J.-F. (1994). Test of homogeneity of binary data with explanatory variables. *Biometrics* **50,** 613–620.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes, Characterization and Convergence.* New York: Wiley.

Fraser, D. W. (1983). Clustering of disease in population units: An exact test and its asymptotic version. *American Journal of Epidemiology* **118,** 732–739.

Haber, M., Longini, I. M., and Cotsonis, G. A. (1988). Models for the statistical analysis of infectious disease data. *Biometrics* **44,** 163–173.

Lefèvre, C. (1990). Stochastic epidemic models for S-I-R infectious diseases: A brief survey of the recent theory. In *Stochastic Processes in Epidemic Theory. Lecture Notes in Biomathematics* **86,** J.-P. Gabriel, C. Lefèvre, and P. Picard (eds), 1–12. Berlin: Springer.

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

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

Schork, N. J. (1994). Sampling guidelines for testing secondary attack rates associated with short-latency infectious diseases. *Statistics in Medicine* **13,** 1563–1573.

Smith, P. G. and Pike, M. C. (1976). Generalisations of two tests for the detection of household aggregation of disease. *Biometrics* **32,** 817–828.

Walter, S. D. (1974). On the detection of household aggregation of disease. *Biometrics* **30,** 525–538.

Received April 1995; *revised January* 1996 *and May* 1996; *accepted July* 1996.

Appendix

In this section we will derive the score statistic (3.1) heuristically. A detailed proof is given in Britton (1996).

First we treat the case where all families are of equal size, $f \geq 2$, which implies that the population size N is equal to nf. For each pair of integers (i, r) such that $0 \leq i \leq i+r \leq f$, we say that a family is in state (i, r) at time t if the family has i infective and r removed individuals. From state (i, r) a family can "jump" to either of two other states: it can jump to state $(i+1, r)$, meaning a susceptible individual in the family was infected, or it can jump to state $(i-1, r+1)$ if an infective individual was removed. Thus, we define two counting processes: $M_{i,r,1}(t)$ is the number of families that have jumped from state (i, r) to $(i+1, r)$, and $M_{i,r,2}(t)$ is the number of (i, r) to $(i-1, r+1)$ jumps. Let $M_{i,r}(t) := M_{i-1,r,1}(t) + M_{i+1,r-1,2}(t) - M_{i,r,1}(t) - M_{i,r,2}(t)$, the number of families in state (i, r), and $I(t) := M_{\cdot,\cdot,1}(t) - M_{\cdot,\cdot,2}(t) := \Sigma_{i,r} M_{i,r,1}(t) - \Sigma_{i,r} M_{i,r,2}(t)$, the number of infectives. From the model description of Section 2.1 it follows that the counting processes $M_{i,r,1}(t)$ and $M_{i,r,2}(t)$ have intensities

$$\begin{aligned} \lambda_{i,r,1}(t) &= M_{i,r}(t-)(f-i-r)\left(\lambda I(t-)/(nf-1) + \delta i/(f-1)\right) \\ \lambda_{i,r,2}(t) &= M_{i,r}(t-)i, \end{aligned} \tag{A.1}$$

respectively. The factor $M_{i,r}(t-)(f-i-r)$ is the number of susceptible individuals in families in state (i, r) just before t and $M_{i,r}(t-)i$ is the corresponding number of infectives. The remaining parts in (A.1) are the rates with which each of these individuals get infected/removed, respectively (see the end of Section 2.1).

Suppose we start with a small initial proportion of infectives, with the remaining part of the population being susceptible (we could also assume that the epidemic is started with a fixed finite number of infective individuals; then the results below are valid conditioned on outcome of a major epidemic). Using theory for Markov processes (cf. Ethier and Kurtz, 1986), it can be shown that, as the number of families n tends to infinity, $\sqrt{n}(n^{-1}\mathbf{M}(\cdot) - \mathbf{m}(\cdot)) \Rightarrow \mathbf{V}(\cdot)$, where $\mathbf{V}(\cdot)$ is a Gaussian Markov vector process. By the notation $\mathbf{M}(\cdot)$ we mean the counting processes $\{M_{i,r,k}(\cdot)\}$ vectorized in some way. The vector of deterministic functions $\mathbf{m}(\cdot)$, vectorized accordingly, is the solution to the set of differential equations defined similarly to (A.1):

$$\dot{m}_{i,r,1}(t) = m_{i,r}(t)(f-i-r)\left(\lambda\eta(t) + \frac{\delta i}{f-1}\right) \quad \text{and} \quad \dot{m}_{i,r,2}(t) = m_{i,r}(t)i, \tag{A.2}$$

where $m_{i,r}(t) := m_{i-1,r,1}(t) + m_{i+1,r-1,2}(t) - m_{i,r,1}(t) - m_{i,r,2}(t)$, and $\eta(t) := m_{\cdot,\cdot,1}(t)/f - m_{\cdot,\cdot,2}(t)/f$. So $m_{i,r}(t)$ should be interpreted as the asymptotic probability of a family being in state (i, r) at t and $\eta(t)$ should be interpreted as the asymptotic probability of being infective at t. For the future, we also define the corresponding probability of being removed as $\rho(t) := m_{\cdot,\cdot,2}(t)/f$. Time derivatives will be denoted by a dot, as in (A.2), and derivatives with respect to δ by a prime.

Let $N_{f,i}$ denote the number of families with i infected individuals at the end of the epidemic and let $\mathbf{N} = (N_{f,0}, \ldots, N_{f,f})^{\mathrm{T}}$. At the end, all infected individuals will be removed, so $N_{f,i} = M_{0,i}(\infty)$ with the earlier notation. Similarly, we let $\mathbf{p}(\delta) = (p_{f,0}(\delta), \ldots, p_{f,f}(\delta))^{\mathrm{T}}$, where $p_{f,i}(\delta)$ denotes the asymptotic probability of ending up with i infected individuals, which for the same reason equals $m_{0,i}(\infty)$. From the weak convergence result for $\mathbf{M}(\cdot)$ mentioned above, the following CLT can be shown: $\sqrt{n}(n^{-1}\mathbf{N} - \mathbf{p}(\delta)) \xrightarrow{D} N(\mathbf{0}, \Sigma(\delta))$. The vector $\mathbf{p}(\delta)$ and the matrix $\Sigma(\delta)$ also depend on the parameter λ, which is suppressed for convenience. The CLT gives the following approximation of the log likelihood $l(\delta) := \log \mathrm{P}_\delta(\mathbf{N} = \mathbf{n})$:

$$l(\delta) \approx c_n(\delta) - \frac{1}{2n}\left(\mathbf{n} - n\mathbf{p}(\delta)\right)^{\mathrm{T}} \Sigma^{-}(\delta)\left(\mathbf{n} - n\mathbf{p}(\delta)\right),$$

where $c_n(\delta)$ is independent of the data and $\Sigma^{-}(\delta)$ is a generalized inverse of the singular matrix $\Sigma(\delta)$. Differentiating both sides with respect to δ and setting $\delta = 0$ gives us the approximation

$$l'(0) \approx c_n'(0) + \mathbf{p}'(0)^{\mathrm{T}}\Sigma^{-}(0)\,(\mathbf{n} - n\mathbf{p}(0)) - \frac{1}{2n}\,(\mathbf{n} - n\mathbf{p}(0))^{\mathrm{T}}\,\Sigma^{-\prime}(0)\,(\mathbf{n} - n\mathbf{p}(0))\,.$$

On the right-hand side, the first term is independent of the data and irrelevant for the construction of a test statistic. The second term is of order $\sqrt{n}$, whereas the third is of order 1, which implies that it can be neglected. Further, in the second term only the part containing $\mathbf{n}$ is relevant because the other is independent of the data. For positive δ, the quantities $\mathbf{p}(\delta)$ and $\Sigma(\delta)$ are only known implicitly. However, when $\delta = 0$, it can be shown that $\Sigma(0) = \mathrm{diag}(\mathbf{p}(0)) - \mathbf{p}(0)\mathbf{p}(0)^{\mathrm{T}}$, $p_{f,i}(0) = \binom{f}{i}p(0)^i(1 - p(0))^{f-i}$. (For general δ, $p(\delta)$ is the asymptotic probability of ever getting infected.) This implies that $\Sigma^{-}(0)$ is the diagonal matrix with elements $1/p_{f,i}(0)$. The test of the hypothesis $\delta = 0$ should hence be based on

$$\mathbf{p}'(0)^{\mathrm{T}}\Sigma^{-}(0)\mathbf{N} = \sum_{i=0}^{f} \frac{p_{f,i}'(0)}{p_{f,i}(0)} N_{f,i}. \tag{A.3}$$

To be able to compute (A.3) it remains to derive $p_{f,i}'(0)$.

The functions $\{m_{i,r,k}(\cdot)\}$ are continuously differentiable, increasing and bounded by 1, so we may change the order of letting $\delta \to 0$ and $t \to \infty$. The same holds for $m_{i,r}(t)$. From now on we indicate the dependence on δ by a superior index. Define $g_0^{(\delta)}(t) := 0$ and, for $k = 1, \ldots, f$, let $g_k^{(\delta)}(t) := \Sigma_{\{i+r \geq k\}}\, m_{i,r}^{(\delta)}(t)$, the asymptotic probability of having k or more infected individuals at t. This implies that $p_{f,i}(\delta) = g_i^{(\delta)}(\infty) - g_{i+1}^{(\delta)}(\infty)$, so $p_{f,i}'(0) = (\partial/\partial\delta)g_i^{(\delta)}(\infty) - g_{i+1}^{(\delta)}(\infty)|_{\delta=0}$. From the defining differential equations of $m_{i,r,k}^{(\delta)}(t)$ in (A.2) we get

$$\dot{g}_k^{(\delta)}(t) = (f - k + 1)\left(\lambda\eta^{(\delta)}(t)\left(g_{k-1}^{(\delta)}(t) - g_k^{(\delta)}(t)\right) + \frac{\delta}{f-1}\sum_{i=1}^{k-1} i m_{i,k-1-i}^{(\delta)}(t)\right). \tag{A.4}$$

The differential equations are intuitive: the function $g_k^{(\delta)}$ increases when susceptible individuals in families with $f - k + 1$ susceptibles become infected. All such susceptibles are exposed to the pressure $\lambda\eta^{(\delta)}(t)$. Susceptibles in families in state $(i, k - 1 - i)$ are exposed to the extra pressure $i\delta/(f - 1)$ from family members.

We use two observations to derive $p_{f,i}'(0)$ from (A.4). The first observation is that we may use $m_{i,k-1-i}^{(\delta)}(t)$ and $m_{i,k-1-i}^{(0)}(t)$ interchangeably in (A.4) because it has δ in front and we are only interested in small δ. The second observation is that, for $\delta = 0$, $m_{i,r}^{(0)}(t) = [f!/(i!r!(f - i - r)!)]\eta^{(0)}(t)^i\rho^{(0)}(t)^r(1 - \eta^{(0)}(t) - \rho^{(0)}(t))^{(f-i-r)}$. This is not hard to show. When $\delta = 0$, the differential system is simplified drastically, but it is also clear from symmetry. From these two observations we obtain the approximation below of the differential equation for $\eta^{(\delta)}(t) + \rho^{(\delta)}(t) = \Sigma_{k=1}^{f}\, g_k^{(\delta)}(t)/f$; the differential equation for $\rho^{(\delta)}(t) = m_{\cdot,\cdot,2}^{(\delta)}(t)/f$ is obtained from (A.2),

$$\dot{\eta}^{(\delta)}(t) + \dot{\rho}^{(\delta)}(t) = \eta^{(\delta)}(t)\left(1 - \eta^{(\delta)}(t) - \rho^{(\delta)}(t)\right)(\lambda + \delta) + o(\delta),$$
$$\dot{\rho}^{(\delta)}(t) = \eta^{(\delta)}(t).$$

Combining these, we see that $\log(1 - \eta^{(\delta)}(t) - \rho^{(\delta)}(t)) = -(\lambda + \delta)\rho^{(\delta)}(t) + o(\delta)$. Further, one can show that $\eta^{(\delta)}(\infty) = 0$, a fact that is intuitively clear because there will be no infective individuals at the end. Thus, $p(\delta) := \rho^{(\delta)}(\infty)$ satisfies

$$1 - p(\delta) = e^{-(\lambda+\delta)p(\delta)} + o(\delta). \tag{A.5}$$

Similarly, from (A.4), (A.5), and the two observations, we can get approximations for suitable linear combinations of $g_k^{(\delta)}$. It turns out that $\phi_i^{(\delta)}(t) := \Sigma_{k=1}^{i}\binom{f-k}{i-k}g_k^{(\delta)}(t)$ are simple to work with. This definition implies that $g_i^{(\delta)}(t) - g_{i+1}^{(\delta)}(t) = \Sigma_{k=0}^{i}\,(-1)^{i-k+1}\binom{f-k}{i-k}\phi_{k+1}^{(\delta)}(t)$, and, because

$p_{f,i}(\delta) = g_i^{(\delta)}(\infty) - g_{i+1}^{(\delta)}(\infty)$, this will allow us to calculate $p'_{f,i}(0)$, the quantities needed to compute (A.3). After some calculations, the result turns out to be

$$p'_{f,i}(0) = p_{f,i}(0)\left(f\left(p(0) - \frac{p'(0)}{1-p(0)}\right) + i\left(\frac{p'(0)}{p(0)(1-p(0))} - 2\right) + \frac{i(i-1)}{(f-1)p(0)}\right).$$

The expression $p'(0)$ may be derived from (A.5), but this will not be necessary because only the third term above will be used.

The above was for the case of equal family sizes. If the population has different family sizes, the same technique can be applied. Let π_f denote the proportion of families that are of size f, $\mu := \Sigma_f f\pi_f$ (the average family size), $\gamma := 1 - \pi_1/\mu$. Then it can be shown that

$$\frac{p'_{1,i}(0)}{p_{1,i}(0)} = 1\left(\gamma p(0) - \frac{p'(0)}{1-p(0)}\right) + i\left(\frac{p'(0)}{p(0)(1-p(0))} - \gamma\right),$$

and for $f \geq 2$

$$\frac{p'_{f,i}(0)}{p_{f,i}(0)} = f\left(\gamma p(0) - \frac{p'(0)}{1-p(0)}\right) + i\left(\frac{p'(0)}{p(0)(1-p(0))} - (1+\gamma)\right) + \frac{i(i-1)}{(f-1)p(0)}.$$

Allowing different family sizes, (A.3) therefore becomes

$$\begin{aligned}\mathbf{p}'(0)^{\mathrm{T}}\Sigma^{-}(0)\mathbf{N} = &\left(\gamma p(0) - \frac{p'(0)}{1-p(0)}\right)\sum_{f=1}^{f_{max}}\sum_{i=0}^{f} fN_{f,i} + N_{1,1} \\ &+ \left(\frac{p'(0)}{p(0)(1-p(0))} - (1+\gamma)\right)\sum_{f=1}^{f_{max}}\sum_{i=0}^{f} iN_{f,i} + \frac{1}{p(0)}\sum_{f=2}^{f_{max}}\sum_{i=0}^{f}\frac{i(i-1)}{f-1}N_{f,i}.\end{aligned}$$

The first sum is the size of the population (a known constant) and the sum in the third term is the number of infected individuals which we condition upon. The test should hence be based on the second and fourth terms, and if we multiply with $p(0)$ and replace $p(0)$ by $\tilde{p}$, the observed proportion infected, this gives us (3.1).