© Board of the Foundation of the Scandinavian Journal of Statistics 1997. Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA. Vol 24: 315–330, 1997

A Test to Detect Within-family Infectivity when the Whole Epidemic Process is Observed

TOM BRITTON

Stockholm University

ABSTRACT. An epidemic model for the spread of an infectious disease in a population of families is considered. The score test of the hypothesis that there is no higher infectivity between family members is constructed under the assumption that the epidemic process is observed continuously up to some time *t*. The score process is a martingale as a function of *t* and by letting the number of families tend to infinity, a central limit theorem for the process can be proved. The central limit theorem not only justifies a normal approximation of the test statistic—it also suggests a smaller variance estimator than expected.

Key words: counting process, epidemic model, information, martingale theory, score process, within-family infectivity

1. Introduction

In this paper we study what we call the standard epidemic for a population of families (SEF), an epidemic model allowing an increased infectivity between individuals of the same family. Often the term household is used instead of family, the essential feature is that the population is separated into many small subgroups. We will focus attention on how to test the hypothesis that there is no increased infectivity rate between family members. If the disease we are interested in is transmitted by close person-to-person contacts this hypothesis of course seems very unlikely. However, for several diseases the mechanism by which the disease is transmitted is not known. Thus, the hypothesis should be thought of as modelling more "long-range" contacts which also approximates indirect transmission such as for example transmission through a common water source or air-borne.

For the case when the observed data consist of the final outcome of the epidemic, i.e. for each family we observe how many were infected and how many were not, Britton (1997) has derived the asymptotic locally most powerful test. If $M_{f,i}$ denotes the number of size f families with i individuals infected at the end of the epidemic, and $\tilde{p}$ is the observed overall proportion infected, then this test statistic is

$$T = \tilde{p}M_{1,1} + \sum_{f=2}^{f\max} \sum_{i=0}^{f} \frac{i(i-1)}{f-1} M_{f,i}. \tag{1.1}$$

The conditional distribution of T, given $\tilde{p}$, is approximately (asymptotically) Gaussian. Under the null hypothesis the mean is $\mu_T = \tilde{p}^2 N$ and the variance is

$$\sigma_T^2 = 2\tilde{p}^2\tilde{p}^2 \textstyle\sum_{f=2}^{f\max} m_f f/(f-1) + \tilde{p}^3\tilde{q}(m_1 - m_1^2/N),$$

where m_f is the number of size f families, $N = \sum_f fm_f$ is the size of the population and $\tilde{q} = 1 - \tilde{p}$.

There are several other tests suggested for the same hypothesis under other similar models; see Britton (1997) for a survey. Like (1.1), all these tests are based on the final outcome of the epidemic.

In this paper we construct a test of the same hypothesis but now assuming that the whole epidemic process is observed up to some time. The test statistic (given in (3.1)), which now is a function of how long we have observed the epidemic, is known as the score process; it is the derivative of the log likelihood with respect to the "within-family" parameter. Even for a population of moderate size, the exact distribution of this process is not known. However, it follows from theory for counting processes that the process is a martingale. When the number of families is increased we may therefore use martingale theory to prove a central limit theorem (CLT) for the score process.

An interesting implication of the CLT is that the variance function of the limit process is strictly less than the limit of the sequence of variance functions. For example, if we observe the whole epidemic until it terminates, it is seen that the variance of the limit variable is only one half the size of the limit of the sequence of variances. Thus, the CLT not only justifies a normal approximation, it also corrects the size of the test one would obtain if the limit of the variance functions was used when normalizing.

Further, the asymptotic results tell us that the amount of information our test is based upon is *not* proportional to the number of infections we observe, but rather information increases linearly with the observation-time of the epidemic process. In other words, infections occurring at "odd" times carry more information.

This is the plan for the rest of the paper: In section 2 we define the model and derive the likelihood for a realization. In section 3 we state our main results concerning the score process and its asymptotic behaviour. In section 4 we interpret the asymptotic results and discuss possible generalizations and extensions. The longer proofs are found in the appendix.

2. The standard epidemic for a population of families

SEF is an extension of the well-known general epidemic model (GE). The first two properties are identical with GE: (i) the length of infectious periods are i.i.d. with an exponential distribution having mean $1/\gamma$, and (ii) during an individual's infectious period, she makes contact with other individuals according to a Poisson process with intensity parameter λ, each contact is made with an individual chosen at random from the whole population (beside herself). In SEF we also have an extra rate to infect family members: during the infectious period she also makes contact according to another Poisson process with intensity δ. For this contact process the contact is made with a family member and each family member has equal probability of being chosen. If a contact is made with an individual not yet infected this individual becomes infected and infective, otherwise nothing happens. All Poisson processes, infectious periods and random selection numbers are defined mutually independent. When the infectious period terminates an individual is considered to be immune, said to be removed, and plays no further role in the epidemic.

SEF is a special case of models described by other authors, for example Ball & Clancy (1993) (who treat a different asymptotic situation, however) and the model in Ball *et al*. (1997) slightly modified. Ball *et al*. (1997) generalize SEF by letting the length of the infectious period follow an arbitrary distribution, still being i.i.d. for different individuals. We are, however, not only interested in the final outcome of the epidemic but the whole epidemic process. To retain the Markov structure of the process we impose the restriction of exponentially distributed infectious periods. The restriction can be relaxed to let the infectious period follow any phase-type distribution, a class of distributions dense in the class *all* distributions on the positive axis; more about this and other generalizations in section 4.

Suppose there are n families, labelled $1, \ldots, n$, in the population and let f_i denote the size of family i. At the start there has to be someone infective for anything to happen, assume for example that one individual in family 1 is infective at $t = 0$. Let $S_i(t)$, $I_i(t)$ and $R_i(t)$ respectively denote the number of susceptible, infective and removed in family i at time t and let $S(t)$, $I(t)$ and $R(t)$ be the corresponding population totals ($S(t) = \sum_i S_i(t)$ etc.). Note that we have the linear relation $S_i(t) + I_i(t) + R_i(t) = f_i$ and hence also $S(t) + I(t) + R(t) = \sum_i f_i =: N$ (the population size). At time instants when an individual changes state (i.e. become infected or removed) we define the individual to be in the *new* state which means that the processes are right continuous. Let $\{\mathscr{F}_t\}_{t\geqslant 0}$ be the natural filtration, $\mathscr{F}_t := \sigma((S_i(s), I_i(s), R_i(s)); 0 \leqslant s \leqslant t, i, \ldots, n)$.

In later sections we will assume n, the number of families, gets large. Whenever this has to be emphasized relevant symbols will have an n attached to them.

If $N_i(t) := I_i(t) + R_i(t) = f_i - S_i(t)$ denotes the number of individuals in family i that have been infected before (or at) t, SEF can be defined using counting processes. The vector $\mathbf{N}(t) = (N_1(t), R_1(t), \ldots, N_n(t), R_n(t))$ is a $2n$-dimensional counting process with $\mathscr{F}_t$-intensities

$$\lambda_i(t) = S_i(t-)\left(\frac{\lambda}{N-1} I(t-) + \frac{\delta}{f_i - 1} I_t(t-)\right) \quad \text{for } N_i(t) \text{ and}$$

$$\gamma_i(t) = \gamma I_i(t-) \quad \text{for } R_i(t). \tag{2.1}$$

This is true because $N_i(t)$ increases by 1 iff a susceptible individual in family i has contact with an infective individual; there are $S_i(t-)$ susceptible individuals in family i and each of them have contact with a given infective individual of a different family at rate $\lambda/(N-1)$ and a given infective individual of the same family at rate $\lambda/(N-1) + \delta/(f_i - 1)$. This explains $\lambda_i(t)$. The intensity $\gamma_i(t) = \gamma I_i(t-)$ follows because $R_i(t)$ increases by 1 iff an infective in family i is removed and each infective is removed at rate γ. The vector $\mathbf{N}(t)$ only jumps one component at a time since these processes can be defined through stochastic time changes of independent Poisson processes, and it is well-known that a finite number of independent Poisson processes have distinct jumps with probability 1. If the family size is 1 ($f_i = 1$), we have 0 in the denominator but then either $S_i(t-) = 0$ or $I_i(t-) = 0$, so formulas will be consistent if we adopt the convention $0/0 = 0$ which is assumed throughout this paper.

Let Q denote the probability measure for which $\mathbf{N}(\cdot)$ is a vector of independent Poisson processes with constant intensity 1. If we observe the epidemic up to time t, it is well known from counting process theory (cf. Andersen *et al.*, 1993) that the likelihood relative to Q is given by

$$\left.\frac{dP}{dQ}\right|_{\mathscr{F}_t} = \exp\left(\sum_{i=1}^{n} \int_0^t \log\lambda_i(s)\, dN_i(s) + \log\gamma_i(s)\, dR_i(s) - (\lambda_i(s) + \gamma_i(s) - 2)\, ds\right). \tag{2.2}$$

Later we want to make inference on the parameter δ; in particular we want to test the hypothesis $\delta = 0$ against the alternative $\delta > 0$. We therefore supress λ and γ in notation and let P_δ denote the probability measure which has δ as "within-family" parameter; for the same reason we write $\lambda_i(\cdot; \delta)$. We base our statistical test on the likelihood ratio $L_t(\delta) = dP_\delta/dP_0|_{\mathscr{F}_t}$. The most powerful test is to reject the hypothesis whenever $L_t(\delta)$, or equivalently $l_t(\delta) = \log L_t(\delta)$, is large. Unfortunately, in our specific case the test depends on δ and no uniformly most powerful test exists. We proceed by maximizing the power for small δ and hence use the score statistic $l'_t(0)$ as our test statistic.

3. Main results

In counting process theory it is known (cf. Andersen *et al.*, 1993) that the derivative of the log likelihood of a counting process, called the score process, defines a local martingale. In the present model this can be checked explicitly; using (2.1) and (2.2) straightforward calculations render

$$l_t'(\delta)=\sum_{i=1}^{n}\left(\int_0^t \frac{\lambda_i'(s;\,\delta)}{\lambda_i(s;\,\delta)}\,dN_i(s)-\lambda_i'(s;\,\delta)\,ds\right)$$
$$=\sum_{i:f_i>1}\int_0^t \frac{I_i(s-)/f_i-1)}{\lambda I(s-)/(N-1)+\delta I_i(s-)/(f_i-1)}(dN_i(s)-\lambda_i(s;\,\delta)\,ds),$$

where $\lambda_i'(s;\,\delta)$ denotes the derivative with respect to δ. Because $\lambda_i(s;\,\delta)$ is the intensity of $N_i(s)$ under P_δ, $l_t'(\delta)$ is seen to be a $(P_\delta, \mathscr{F}_t)$-local martingale, it is even a martingale since $l_t'(\delta)$ is bounded for fixed n. We have thus shown the following.

Proposition

If we observe SEF up to t, the locally most powerful test (the score test) of the hypothesis $\delta=0$ *against alternatives* $\delta>0$ *is based on*

$$Y(t):=\lambda l_t'(0)=\sum_{i:f_i>1}\int_0^t \frac{I_i(s-)/(f_i-1)}{I(s-)/(N-1)}(dN_i(s)-\lambda_i(s;\,0)\,ds). \tag{3.1}$$

If we look at $Y(t)$, it makes sense as a test of $\delta=0$. If many infections occur in families with a relatively high proportion of infectives our statistic will become large and this also speaks in favour of the alternative.

This kind of test statistic is only useful if its distribution under the null hypothesis, P_0, is known. Below we prove a CLT for the process $Y(t)$ rescaled in time and size. For a population with many families we may use this result to approximate the distribution of $Y(t)$.

From now on we will only use P_0 in this paper. Under P_0 SEF reduces to GE (the general epidemic model): the family structure is irrelevant when $\delta=0$ since there is no extra within-family infectivity. We may thus use known results from this model when proving a CLT. One such result (e.g. Martin-Löf, 1986) is that as the population size grows, the final proportion infected is concentrated around the solutions to the equation $x=1-\exp(-\lambda x/\gamma)$. If $\lambda\leqslant\gamma$, 0 is the only solution. In fact, when this is the case, the number of infected individuals is known to be bounded in probability as the population size tends to infinity. If $\lambda>\gamma$, there are two solutions, 0 and a second solution

$$\pi\in(0,\,1)\quad\text{is the positive solution to the equation}\quad x=1-\exp(-\lambda x/\gamma). \tag{3.2}$$

We call it a *major epidemic* if we end up close to π. More precisely, if $R^{(n)}(\infty)$ is the final number of removed, i.e. the total number of individuals ultimately infected by the epidemic, we say that a major epidemic has occurred if $R^{(n)}(\infty)/N^{(n)}>\pi/2$ ($N^{(n)}$ is the population size). It is also known for GE that the number of infections in any finite time interval $[0,\,t]$ is bounded in probability as the population size grows (e.g. Ball, 1995). This result is true for any choice of λ and γ. In the case of a major epidemic the duration of the epidemic is known to be of order $\log n$.

Heuristically, many random events must be under consideration for a CLT to have a chance to be valid. For this reason and because of the properties of GE mentioned above, we assume that $\lambda>\gamma$ and that the time for which the epidemic is observed grows like $\log n$. The correct, somewhat unusual, scaling of our martingale turns out to be

$$M^{(n)}(t):=\frac{Y^{(n)}(t\log n)}{\sqrt{n}\log n}\frac{N^{(n)}}{N^{(n)}-1}$$

$$= \frac{N^{(n)}}{n}\sqrt{\frac{n}{\log n}} \sum_{i: f_i > 1} \int_0^{t \log n} \frac{I_i^{(n)}(s-)/(f_i - 1)}{I^{(n)}(s-)} (dN_i^{(n)}(s) - \lambda_i^{(n)}(s;\, 0)\, ds) \tag{3.3}$$

which is a martingale with respect to $P_0^{(n)}$ and the filtration $\tilde{\mathscr{F}}_t^{(n)} := \mathscr{F}_{t \log n}^{(n)}$. The factor $N^{(n)}/(N^{(n)} - 1)$ is asymptotically irrelevant but is there to simplify notation in what follows. Note that the integration stops if $I^{(n)}(s-) = 0$ because then $I_i^{(n)}(t-) = 0$, $t \geqslant s$ for each i so the integrand is $0/0$ which we defined as 0 in section 2.

As the population grows we have to make some regularity assumptions on the family-size frequencies. We make things easy for us by assuming that $m_f^{(n)}$, the number of families of size f, satisfies $m_f^{(n)}/n = p_f$ independent of n and that $p_f = 0$ for $f > f_{\max}$, so $f_{\max}$ is the largest family size. Let $\mu = \sum_f f p_f = N^{(n)}/n$ denote the average family size, the first factor in the second row of (3.3) is then μ.

The optional and predictable variation processes of $M^{(n)}$ are given by

$$[M^{(n)}](t) = \frac{\mu^2 n}{\log n} \sum_{i: f_i > 1} \int_0^{t \log n} \left(\frac{I_i^{(n)}(s-)}{(f_i - 1) I^{(n)}(s-)} \right)^2 dN_i^{(n)}(s) \tag{3.4}$$

$$\langle M^{(n)} \rangle (t) = \frac{\mu^2 n}{\log n} \sum_{i: f_i > 1} \int_0^{t \log n} \left(\frac{I_i^{(n)}(s)}{(f_i - 1) I^{(n)}(s)} \right)^2 \lambda_i^{(n)}(s;\, 0)\, ds$$

$$\approx \frac{\lambda \mu}{\log n} \int_0^{t \log n} \frac{\sum_{i: f_i > 1} I_i^{(n)}(s)^2 S_i^{(n)}(s)/(f_i - 1)^2}{I^{(n)}(s)}\, ds \leqslant \lambda \mu t. \tag{3.5}$$

The "$\approx$" in (3.5) would be "$=$" if the right-hand side contained the factor $N/(N - 1)$. The last inequality is true because $I_i^{(n)}(s)^2 S_i^{(n)}(s) \leqslant I_i^{(n)}(s)(f_i - 1)^2$ so the sum in the numerator is less than the denominator. (See for example Andersen *et al.* (1993) for more properties on martingales and their optional and predictable variation processes.) This implies bounded second moments of $M^{(n)}(t)$ for each t

$$\mathrm{var}\,(M^{(n)}(t)) = E([M^{(n)}](t)) = E(\langle M^{(n)} \rangle (t)) \leqslant \lambda \mu t. \tag{3.6}$$

We will now split up $M^{(n)}$ into three martingales, the first containing all very large jumps of $M^{(n)}$, the second containing the remaining jumps that are not negligible and the third containing the rest. The size of a jump $M^{(n)}$ may do at time s is smaller than, and of the same order as, $\mu\sqrt{n}/(\sqrt{\log n}\, I^{(n)}(s-))$. Hence we define the events $A_k^{(n)}(s)$, $k = 1, 2, 3$, by

$$A_1^{(n)}(s) := \left\{ 1 \leqslant I^{(n)}(s-) < \sqrt{\frac{n}{\log n}} \frac{1}{g(n)} \right\}$$

$$A_2^{(n)}(s) := \left\{ \sqrt{\frac{n}{\log n}} \frac{1}{g(n)} \leqslant I^{(n)}(s-) < \sqrt{\frac{n}{\log n}} g(n) \right\}$$

$$A_3^{(n)}(s) := \left\{ \sqrt{\frac{n}{\log n}} g(n) \leqslant I^{(n)}(s-) \right\},$$

where $g(n)$ is a function tending to infinity slow enough ($g(n) = (\log n)^{1/4}$ will do). Let $1_{A_k^{(n)}(s)}$ denote the corresponding indicator functions and define from them three martingales. For $k = 1, 2, 3$, let

$$M_k^{(n)}(t) := \mu\sqrt{\frac{n}{\log n}} \sum_{i:f_i>1} \int_0^{t\log n} \frac{1_{A_k^{(n)}(s)} I_i^{(n)}(s-)}{(f_i-1)I^{(n)}(s-)} (dN_i^{(n)}(s) - \lambda_i^{(n)}(s;\,0)\,ds).$$

The following three properties are easy to check: (i) $M_1^{(n)}$, $M_2^{(n)}$ and $M_3^{(n)}$ are martingales. (ii) $M^{(n)} = M_1^{(n)} + M_2^{(n)} + M_3^{(n)}$. (iii) The largest possible jump in $M_3^{(n)}$ tends to 0 as $n \to \infty$.

We will show that $M^{(n)}$ converges to a Gaussian process but this will be the case only if we have a major epidemic. This means we would like to condition on the event that there is a major epidemic but then $M^{(n)}$ is no longer a martingale. However, we can modify $M^{(n)}$ to let it start at a stopping time defined as the first time when a fairly large number, say $\log n$, of individuals have been infected and this modified process is a martingale with respect to the naturally modified filtration (see sect. II.4.4 in Andersen *et al.* (1993) for details of such martingales). It is known (Martin-Löf, 1986) that the event that $\log n$ individuals actually will get infected, asymptotically coincides with the event that there is a major epidemic. Further, the modified process and $M^{(n)}$ will be asymptotically identical because whatever happens before the stopping time is asymptotically negligible, as is the time until the stopping time occurs (on the $\log n$-scale). This is how the conditioning event in the results below should be treated in order apply theory for martingales.

Before showing weak convergence of $M^{(n)}$ we state some preliminary results, the corresponding proofs are found in the appendix.

Lemma 1
$M_1^{(n)} \Rightarrow 0$ (weak convergence to the constant 0-process).

Lemma 2

$$\langle M_2^{(n)}\rangle(t) \overset{P_0^{(n)}}{\to} 0 \quad \forall t \geqslant 0.$$

Lemma 3
Given that there is a major epidemic,

$$\langle M_3^{(n)}\rangle(t) \overset{P_0^{(n)}}{\to} v(t) \quad \forall t \geqslant 0,$$

where

$$v(t) = \alpha((t-t_1)^+ \wedge t_1) + (1-\pi)\alpha((t-2t_1)^+ \wedge t_2),$$

$\alpha = \lambda\sum_{f=2}^{f_{\max}} fp_f/(f-1)$, $2t_1 = 1/(\lambda-\gamma)$ *and* $2t_2 = 1/(\gamma-(1-\pi)\lambda)$.

Remark. Hence $v(t)$ equals 0 until t_1, it then starts increasing linearly with slope α until $t = 2t_1$ when its linear growth decreases to $(1-\pi)\alpha$, after $t = 2t_1 + t_2$ it remains constant. See Fig. 1 below.

Once the lemmas above have been proven it is straightforward to show our main result.

Theorem
Let $M^{(n)}$ be the $(P_0^{(n)}, \tilde{\mathscr{F}}_t^{(n)})$ martingale defined in (3.3). *Then, given that there is a major epidemic,*

$$M^{(n)} \Rightarrow M \qquad \text{as } n \to \infty,$$

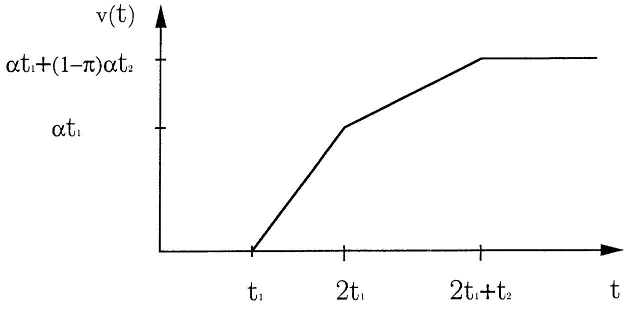

Fig. 1.

where M is a Gaussian martingale with independent increments such that $[M](t) = \langle M\rangle(t) = v(t)$; $v(t)$ *is defined in lemma 3 and plotted in Fig. 1.*

Proof. Lemma 2, 3 and the discussion before lemma 1 together imply that $M_2^{(n)} + M_3^{(n)} \Rightarrow M$ by Rebolledo's theorem for local martingales (th. II.5.1 in Andersen *et al.*, 1993, or Rebolledo, 1980). The martingale $M_1^{(n)}$ has to be treated separately because $\langle M_1^{(n)}\rangle(t) \xrightarrow{p} h(t)$, a strictly positive deterministic function, so Rebolledo's theorem would not go through on $M^{(n)}$ itself (the reason why not is that, even though the "large-jumps-part" of $M^{(n)}$, $M_1^{(n)}$, converges in probability to the "0-process", its variation process $\langle M_1^{(n)}\rangle$ does not converge to 0 which is needed in Rebolledo's theorem). However, by lemma 1 we may apply Slutsky's theorem on $M_1^{(n)}$ and $M_2^{(n)} + M_3^{(n)}$ to conclude the statement of the theorem.

The theorem justifies a normal approximation to the distribution of $Y(t)$ in (3.1). Unfortunately, $Y(t)$ contains the nuisance parameter λ and the variance function $v(t)$ is a function of λ and γ. It is natural to replace these parameters by their ML-estimators but it is not obvious that the same asymptotic results are still valid. This turns out to be the case however, and we state the result as a corollary. The proof is found in the appendix.

Let $\hat{M}^{(n)}$ be like $M^{(n)}$ with the parameter λ replaced by its ML-estimator under the model with $\delta = 0$, that is

$$\hat{M}^{(n)}(t) := \mu\sqrt{\frac{n}{\log n}} \sum_{i:f_i>1} \int_0^{t\log n} \frac{I_i^{(n)}(s-)}{(f_i-1)I^{(n)}(s-)} \left(dN_i^{(n)}(s) - \hat{\lambda}_t^{(n)} S_i^{(n)}(s) \frac{I^{(n)}(s)}{N^{(n)}-1}\, ds \right)$$

where

$$\hat{\lambda}_t^{(n)} = N^{(n)}(t\log n) \Big/ \int_0^{t\log n} I^{(n)}(s) \frac{S^{(n)}(s)}{N^{(n)}-1}\, ds.$$

Corollary

Given that there is a major epidemic, $\hat{M}^{(n)} \Rightarrow M$ *as* $n \to \infty$, *where M is as in the theorem. Further,* $v(t)$, *the variance function of M, may be consistently estimated by* $\hat{v}^{(n)}(t)$ *which is defined like* $v(t)$ *only with* λ, γ *and* π *replaced by* $\hat{\lambda}_t^{(n)}$, $\hat{\gamma}_t^{(n)} = R^{(n)}(t\log n)/\int_0^{t\log n} I^{(n)}(s-)\,ds$ *and* $\hat{\pi}^{(n)}$ *which is the solution to* (3.2) *with* $\hat{\lambda}_t^{(n)}$ *and* $\hat{\gamma}_t^{(n)}$.

If we observe a SEF up to some time t and want to test the hypothesis $\delta = 0$ the corollary implies that we shall use the following.

Test procedure

Reject the hypothesis $\delta = 0$ *if* $\hat{Y}(t)$ *is large compared with the normal distribution with mean 0 and variance* $(N - 1/N)^2 \hat{v}(t/\log n) n \log n$. *Here* $\hat{Y}(t)$ *is the statistic obtained by replacing* λ *with* $\hat{\lambda}_t$ *in* $Y(t)$, *and as before N is the population size and n denotes the number of families.*

4. Discussion

We have shown that the normed score process, $M^{(n)}$, converges to a Gaussian process with variance function $v(t)$ (see Fig. 1). This variance function is not the same as the limit of $\text{var}(M^{(n)}(t))$. It is straightforward from lemma 5 to show that $\text{var}(M^{(n)}(t))$ converges to a function which increases with the same slopes as $v(t)$ but now on the double sized intervals $(0, 2t_1)$ and $(2t_1, 2t_1 + 2t_2)$ respectively. In particular, if we observe the whole epidemic the variance of the normed limit variable is

$$v(\infty) = v(2t_1 + t_2) = \alpha(t_1 + t_2(1 - \pi)) = \frac{\gamma\pi \sum_{f=2}^{f\max} fp_f/(f-1)}{2\lambda(\lambda - \gamma)(\gamma - (1-\pi)\lambda)}$$

which is only half the limit of $\text{var}(M^{(n)}(\infty))$. The CLT thus tells us that the more crude method to normalize with the limiting standard deviation gives the wrong size of the test.

In the theory of inference, $v(t)$ is known as the expected information. It increases linearly which means that the information is proportional to the time we observe the epidemic, a result which might seem obvious at first. On the other hand intuition tells us that the information should be proportional to the number of infections we observe but this is not true. Almost all infections, a proportion tending to 1, take place arbitrary close to $2t_1 = 1/(\lambda - \gamma)$ on the log time scale; this follows from (A.3). So, loosely speaking, if we observe the epidemic during a time interval of fixed length l (on the log-time scale) we will not have more information to answer the hypothesis if this interval contains the point $2t_1$ and we observe almost all infections, than if the interval does not contain $2t_1$ and we would observe an ever decreasing fraction of all infections. Infections taking place when there are not many infectives in the population hence carry much more information than infections occurring when many individuals are infective. This can be understood because if a susceptible person is infected when any of her family members is infective but very few in the whole population are infective we would be almost certain that there was an increased within-family infectivity. On the other hand if this happened when a positive fraction of the whole population were infective we would not be so sure that the infection was caused by the family member.

In applications the type of data considered in this paper is less common than data consisting of the final outcome of an epidemic. A motivation for the present work is therefore to see how much we would gain by collecting the present, more informative, data. That is, how much power do we gain by collecting this data? It can be seen that we may detect alternatives of order $1/\sqrt{n \log n}$ for the present data whereas when we observe only the final outcome, the alternatives have to be of order $1/\sqrt{n}$ or larger to be detected. This means that we may detect alternatives that are closer to the null hypothesis if we observe the whole epidemic process. However, since the difference is only of order $\sqrt{\log n}$, the gain is moderate and has to be compared with the extra cost of collecting this type of data.

SEF has some properties which might not be very realistic in applications, for example: exponentially distributed infectious periods; that the extra intensity of making contact with family members, δ, is independent of the family size; and the assumption that all individuals behave likewise in terms of getting infected and spreading the disease further.

The assumption of exponentially distributed infectious periods can be relaxed. If instead we assume that these periods follow any phase-type distribution then the score test (3.1) remains unchanged. (A phase-type distribution is defined as the time to absorption in a continuous-time finite-state Markov chain with one absorbing state. This class of distributions is dense in the class of *all* distributions on the positive real line, cf. Asmussen (1987). The likelihood will be different from that in (2.2) but terms coming from the phase-type distribution will cancel when forming the likelihood ratio just like $\gamma_i(\cdot)$ disappeared. The CLT will not be the same in that the variance function $v(t)$ will be different.

The model was defined so that individuals make contact with other family members with increased rate δ. This means that a given infective in a size f family makes contact with a given susceptible of the same family with the additional rate $\delta/(f-1)$. A more general model is of course to let these contact rates be of the form $\delta \times g(f)$ where $g(\cdot)$ is an arbitrary non-negative function. All results in the present paper still hold for this generalized model, just replace $1/(f-1)$ by $g(f)$ everywhere.

Another generalization is to allow individuals to behave differently, that is to have a multi-type population. The different behaviour could be a combination of variable susceptibility, variable infectivity and variable distribution of the infectious period. Although notationally heavier and technically more involved, several of the results in the present paper are straightforward to extend to such models.

Instead of testing for within-family infectivity one can extend the present model to test the relevance of an arbitrary pre-specified contact structure. We would then have a matrix $C = \{c_{j,k}\}$ where $c_{j,k}$ indicates how "close" individual j is to k. A natural extension of the present model then says that during js infectious period she contacts k with the intensity $\lambda/(N-1) + \delta c_{j,k}$. Just like in the family case the null hypothesis is $\delta = 0$. The locally most powerful test for this extended model should be based on

$$Y(t) := \lambda l_t'(0) = \sum_{j=1}^{N} \int_0^t \frac{\sum_k I_k(s-)c_{k,j}}{I(s-)/(N-1)} \left(dN_j(s) - \lambda S_j(s) \frac{I(s-)}{N-1}\, ds \right)$$

which should be compared with (3.1), ($I_k(s)$ is now the indicator for the event that individual k is infective at s and likewise for $S_k(s)$). To obtain a weak convergence result for this process one has to assume some properties on the matrix C but this is beyond the scope of this paper.

In survival analysis similar hypothesis and corresponding test have been derived by Commenges & Andersen (1995). In such models families may be heterogeneous in that each family has a random parameter affecting the hazard rates for individuals belonging to the family. However, given these random parameters, individuals behave independently of each other whereas in the present paper individuals of the same family are genuinely dependent: if someone of your family gets infected this suddenly increases your risk of getting infected.

Acknowledgements

I wish to thank my supervisor Åke Svensson for valuable suggestions and Louise af Klintberg for comments on the manuscript. Thanks also to two referees for careful reading and comments which improved an earlier version of this paper. Financial support from The Bank of Sweden Tercentenary Foundation is gratefully acknowledged.

References

Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical models based on counting processes*. Springer, Berlin.

Asmussen, S. (1987). *Applied probability and queues*. Wiley, Chichester.

Ball, F. (1995). Coupling methods in epidemic theory. In: *Epidemic models: their structure and relation to data* (ed. by Denis Mollison) 34–52. Cambridge University Press, Cambridge.

Ball, F. & Clancy, D. (1993). The final size and severity of a generalised stochastic multitype epidemic model, *Adv. Appl. Probab.* **25**, 721–736.

Ball, F., Mollison, D. & Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Probab.* (in press).

Barbour, A. D. (1975). The duration of the closed stochastic epidemic. *Biometrika* **62**, 477–482.

Britton, T. (1996). A test to detect within-family infectivity when the whole epidemic process is observed. In Epidemics with heterogeneous mixing: stochastic models and statistical tests. PhD thesis, Stockholm University.

Britton, T. (1997). Tests to detect clustering of infected individuals within families, *Biometrics* **53**, (to appear).

Commenges, D. & Andersen, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Anal.* **1**, 145–156.

Martin-Löf, A. (1986). Symmetric sampling procedures, general epidemic processes and their threshold limit theorems, *J. Appl. Probab.* **23**, 265–282.

Rebolledo, R. (1980). Central limit theorems for local martingales. *Z. Wahrsch. Verw. Gebiete*. **51**, 269–286.

Svensson, Å. (1995). On the asymptotic size and duration of a class of epidemic models. *J. Appl. Probab.* **32**, 11–24.

Received November 1995, in final form June 1996

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

Appendix

Let $\epsilon > 0$, assume $\epsilon < \pi/2$ but for the moment otherwise arbitrary (π was defined in (3.2)). Define three stopping times

$$\eta_n := \inf\{t > 0;\ N^{(n)}(t) \geqslant \epsilon\mu n\}$$
$$\rho_n := \inf\{t > 0;\ N^{(n)}(t) \geqslant (\pi - \epsilon)\mu n\}$$
$$\sigma_n := \inf\{t > 0;\ I^{(n)}(t) = 0\}.$$

Remember that μn is the population size and $N^{(n)}(t) := I^{(n)}(t) + R^{(n)}(t)$ is the number of individuals that have been infected by time t. The first two stopping times depend on ϵ which is suppressed in notation. The results below hold independently of ϵ, which is why we may choose ϵ suitably small later.

The following four properties can be deduced from Barbour (1975). In Svensson (1995) they are stated formally for a more general model.

In case of a major epidemic:

$$\eta_n/\log n \overset{P_0^{(n)}}{\to} 1/(\lambda - \gamma) =: 2t_1, \tag{A.1}$$

$$(\sigma_n - \rho_n)/\log n \overset{P_0^{(n)}}{\to} 1/(\gamma - \lambda(1 - \pi)) =: 2t_2, \tag{A.2}$$

$$\rho_n - \eta_n \quad \text{remains bounded in probability, and} \tag{A.3}$$

$$\inf_{\{t;\eta_n \leqslant t \leqslant \rho_n\}} I^{(n)}(t)/\mu n \quad \text{is bounded away from 0 in probability.} \tag{A.4}$$

What these results heuristically say is that, on the log-time scale, more or less everyone is susceptible before $2t_1$ and after this time a proportion π are removed and the rest remain susceptible. Equations (A.1)–(A.4) will be used throughout this section.

Before proving lemmas 1–3 we need a lemma.

Lemma 4

Let $0 < \beta < 1$ *and* $\epsilon > 0$. *In case of a major epidemic*

$$\sup_{\{u; u < \beta 2t_1 - \epsilon\}} I^{(n)}(u \log n)/n^{\beta} \overset{P_0^{(n)}}{\to} 0,$$

$$\inf_{\{u; \beta 2t_1 + \epsilon < u < 2t_1 + (1-\beta)2t_2 - \epsilon\}} I^{(n)}(u \log n)/n^{\beta} \overset{P_0^{(n)}}{\to} \infty,$$

$$\sup_{\{u; u > 2t_1 + (1-\beta)2t_2 + \epsilon\}} I^{(n)}(u \log n)/n^{\beta} \overset{P_0^{(n)}}{\to} 0.$$

The idea behind the proof is that, before $2t_1 \log n$ we can approximate the number of infectives by a birth and death process with positive drift, and after $2t_1 \log n$ we approximate the number of infectives with another birth and death process, now with negative drift and a large initial population. Consequently, we apply results for birth and death processes which together with (A.1)–(A.4) gives the desired result. A detailed proof is found in Britton (1996).

We now prove the three lemmas which are used in the proof of the theorem which is the main result of this paper. The first two are straightforward but the third is a bit more involved although not deep.

Proof of lemma 1. We separate $M_1^{(n)}$ into the integrated counting process and its compensator

$$V_1^{(n)}(t) := \mu \sqrt{\frac{n}{\log n}} \sum_{i: f_i > 1} \int_0^{t \log n} 1_{A_1^{(n)}(s)} \frac{I_i^{(n)}(s-)}{(f_i - 1) I^{(n)}(s-)} \, dN_i^{(n)}(s)$$

$$V_2^{(n)}(t) := \mu \sqrt{\frac{n}{\log n}} \sum_{i: f_i > 1} \int_0^{t \log n} 1_{A_1^{(n)}(s)} \frac{I_i^{(n)}(s-)}{(f_i - 1) I^{(n)}(s-)} \lambda_i^{(n)}(s; 0) \, ds$$

and show that $V_i^{(n)} \Rightarrow 0$, $i = 1, 2$, which will prove the lemma since $M_1^{(n)} = V_1^{(n)} - V_2^{(n)}$. $V_1^{(n)}$ and $V_2^{(n)}$ are non-decreasing in t, their maxima are obtained at $\sigma_n/\log n$ and by (A.1)–(A.3) it follows that

$$\sigma_n/\log n \overset{P_0^{(n)}}{\to} 2(t_1 + t_2).$$

Thus, it suffices to show that

$$V_i^{(n)}(t') \overset{P_0^{(n)}}{\to} 0, \quad i = 1, 2,$$

where $t' > 2(t_1 + t_2)$.

We start with $V_1^{(n)}$. By (3.6),

$$\lambda \mu t' \geqslant E([M^{(n)}](t')) \geqslant E([M_1^{(n)}](t')) = E([M_1^{(n)}](t') | [M_1^{(n)}](t') > 0) P([M_1^{(n)}](t') > 0).$$

The process $[M_1^{(n)}](t')$ is just like (3.4) with $1_{A_1^{(n)}(s)}$ in the integrand. We condition on

$[M_1^{(n)}](t') > 0$ which implies that we must have a jump when $1_{A_1^{(n)}(s)} = 1$. So, by the definition of $A_1^{(n)}(s)$

$$E([M_1^{(n)}](t') | [M_1^{(n)}](t') > 0) > 0) \geqslant \frac{\mu^2 n}{\log n} \frac{1}{(f_{\max} - 1)^2} \frac{g^2(n) \log n}{n} \to \infty,$$

from which we conclude that $P([M_1^{(n)}](t') > 0) \to 0$. But $V_1^{(n)}(t') = 0$ whenever

$$[M_1^{(n)}](t') = 0, \text{ so } V_1^{(n)}(t') \overset{P_0^{(n)}}{\to} 0.$$

We now show that $V_2^{(n)}(t') \overset{P_0^{(n)}}{\to} 0$, the convergence actually holds a.s. (the "$\approx$" below would be an equality if that factor $N/(N-1)$ was added to the right hand side—the same holds in the next proof)

$$\begin{aligned} V_2^{(n)}(t') &\approx \frac{\lambda}{\sqrt{n \log n}} \sum_{i: f_i > 1} \int_0^{t' \log n} 1_{A_1^{(n)}(s)} \frac{I_i^{(n)}(s) S_i^{(n)}(s)}{(f_i - 1)} \, ds \\ &\leqslant \frac{\lambda}{\sqrt{n \log n}} \int_0^{t' \log n} 1_{A_1^{(n)}(s)} I^{(n)}(s) \, \mathrm{d}s \\ &\leqslant \frac{\lambda}{\sqrt{n \log n}} \int_0^{t' \log n} \sqrt{\frac{n}{\log n}} \frac{1}{g(n)} \, ds = \frac{\lambda t'}{g(n)} \to 0. \end{aligned}$$

Proof of lemma 2. Similarly to (3.5),

$$\begin{aligned} \langle M_2^{(n)} \rangle (t) &= \frac{\mu^2 \eta}{\log n} \sum_{i: f_i > 1} \int_0^{t \log n} 1_{A_2^{(n)}(s)} \left(\frac{I_i^{(n)}(s)}{(f_i - 1) I^{(n)}(s)} \right)^2 \lambda_i^{(n)}(s; 0) \, ds \\ &\approx \frac{\mu \lambda}{\log n} \int_0^{t \log n} 1_{A_2^{(n)}(s)} \frac{\sum_{i: f_i > 1} I_i^{(n)}(s)^2 S_i^{(n)} / (f_i - 1)^2}{I^{(n)}(s)} \, ds \\ &\leqslant \frac{\mu \lambda}{\log n} \int_0^{t \log n} 1_{A_2^{(n)}(s)} \, ds = \mu \lambda \int_0^t 1_{\left\{ \sqrt{\frac{n}{\log n} \frac{1}{g(n)}} \leqslant I^{(n)}(u \log n) < \sqrt{\frac{n}{\log n}} g(n) \right\}} \, du. \end{aligned}$$

Apply lemma 4 with $\beta = 1/2 \pm \epsilon$, respectively, for arbitrary small ϵ, to conclude that this integral converges to 0 in probability.

In order to prove lemma 3 we need another lemma (lemma 5 below). The proof of this lemma contains basic combinatorics which may be used because, conditioned on the total numbers of susceptibles and infectives at some time t, each configuration with the right marginals has equal probability under P_0.

We introduce some more notation:

$$N_{f,s,i}^{(n)}(t) := \textstyle\sum_{j=1}^n 1_{\{f_j = f, S_j^{(n)}(t) = s, I_j^{(n)}(t) = i\}}$$

is the number of size f families and s susceptible and i infective individuals at t. Further, let $\mathscr{G}^{(n)} := \sigma(S^{(n)}(t), I^{(n)}(t); t \geqslant 0)$ be the σ-algebra generated by the whole epidemic process if we were not observing the epidemic on family level, so $\mathscr{G}^{(n)} \subset \mathscr{F}^{(n)} := \sigma(\mathscr{F}_t^{(n)}; t \geqslant 0)$.

Whenever possible we will drop n in our notations. Define $\tau_0 = 0$ and let τ_k be the time at which the kth "jump" (either infection or removal) of the process was made. Define the embedded discrete-time epidemic process by $\tilde{S}(k) = S(\tau_k)$, $\tilde{I}(k) = I(\tau_k)$ and $\tilde{R}(k) = R(\tau_k) = \mu n - S(\tau_k) - I(\tau_k)$. Conditioned on $\mathscr{G}^{(n)}$ all these variables are constants.

Below we write

$$\begin{pmatrix} & k & \\ i & & j \end{pmatrix}$$

for the trinomial coefficient $k!/i!j!(k-i-j)!$, and $k_{(l)}$ for $k(k-1)\cdots(k-k+1)$.

Lemma 5
For $1 \leqslant i \leqslant i+s \leqslant f,\ t \geqslant 0$ *and in case of a major epidemic,*

$$\frac{1}{\log n}\int_0^{t\log n} \frac{N^{(n)}_{f,s,i}(u)}{I^{(n)}(u)}\,du \overset{P_0^{(n)}}{\rightarrow} h_{f,s,i}(t)$$

where

$$h_{f,s,i}(t) = 1_{\{i\geqslant 1\}}\begin{pmatrix} & f & \\ i & & s \end{pmatrix}\frac{p_f}{\mu}(1_{\{s=f-1\}}(t\wedge 2t_1) + ((t-2t_1)^+ \wedge 2t_2)(1-\pi)^s\pi^{f-s-1})$$

Proof. Let i, s, f and t be fixed satisfying the assumptions of the lemma, let $r = f-s-i$ and write $X^{(n)}$ for the left-hand side in the lemma. Throughout the proof it is essential to keep in mind that $i \geqslant 1$. We will show that

$$E(X^{(n)}|\mathscr{G}^{(n)}) \overset{P_0^{(n)}}{\rightarrow} h_{f,s,i}(t)$$

and

$$\operatorname{var}(X^{(n)}|\mathscr{G}^{(n)}) \overset{P_0^{(n)}}{\rightarrow} 0.$$

Together with the observation that $0 \leqslant X^{(n)} \leqslant t$, which follows since the numerator in the integral is less than the denominator, this will prove the lemma. In the proof c is a generic constant independent of n and the realisation of the epidemic.

We start with the conditional expectation. Let

$$p_k(s,\ i) = \begin{pmatrix} & f & \\ i & & s \end{pmatrix}\frac{\tilde{I}(k)_{(i)}\tilde{S}(k)_{(s)}\tilde{R}(k)_{(r)}}{(\tilde{I}(k)+\tilde{S}(k)+\tilde{R}(k))_{(f)}} \leqslant c\left(\frac{\tilde{I}(k)}{n}\right)^i, \tag{A.5}$$

the last inequality is true since $\tilde{I}(k)+\tilde{S}(k)+\tilde{R}(k) = \mu n$. Then $p_k(s,\ i)$ is the $\mathscr{G}^{(n)}$-conditional probability that a size f family has s susceptible, i infective and $r = f-s-i$ removed individuals at τ_k. It follows that $E(N_{f,s,i}(\tau_k)|\mathscr{G}^{(n)}) = m_f p_k(s,\ i)$. This implies

$$\begin{aligned} E(X^{(n)}|\mathscr{G}^{(n)}) &= \int_0^t \frac{E(N_{f,s,i}(u\log n)|\mathscr{G}^{(n)})}{I(u\log n)}\,du \\ &= \begin{pmatrix} & f & \\ i & & s \end{pmatrix}\frac{p_f}{\mu}\int_0^t \frac{S(u\log n)_{(s)}(I(u\log n)-1)_{(i-1)}R(u\log n)_{(r)}}{(\mu n-1)_{(f-1)}}\,du. \end{aligned}$$

The last integral converges to $h_{f,s,i}(t)$ in probability because of lemma 4 and (A.1)–(A.4). The key to this is that, for $u < 2t_1$: $S(u\log n)/\mu n \approx 1$ and $I(u\log n)/\mu n \approx R(u\log n)/\mu n \approx 0$, and for $2t_1 < u < 2(t_1+t_2)$: $S(u\log n)/\mu n \approx 1-\pi$, $I(u\log n)/\mu n \approx 0$ and $R(u\log n)/\mu n \approx \pi$.

It remains to show that the conditional variance converges to 0 in probability. When we condition on $\mathscr{G}^{(n)}$ we can split up the integral in $X^{(n)}$ into separate parts corresponding to intervals $[\tau_k,\ \tau_{k+1})$ and the integrand is constant on each such interval. Let $\kappa = \max\{k;\ \tau_k \leqslant t\}$ denote the number of jumps in $[0,\ t]$, and define $\tau_{\kappa+1} = t$. Then

$$\operatorname{var}(X^{(n)}|\mathscr{G}^{(n)}) = \frac{1}{(\log n)^2}\sum_{k=0}^{\kappa}\left(\frac{\tau_{k+1}-\tau_k}{\tilde{I}(k)}\right)^2 \operatorname{var}(N_{f,s,i}(\tau_k)|\mathscr{G}^{(n)})$$
$$+\frac{2}{(\log n)^2}\sum_{0\leqslant j<k\leqslant\kappa}\frac{(\tau_{k+1}-\tau_k)(\tau_{j+1}-\tau_j)}{\tilde{I}(j)\tilde{I}(k)}\operatorname{cov}(N_{f,s,i}(\tau_j),\, N_{f,s,i}(\tau_k)|\mathscr{G}^{(n)}). \qquad \text{(A.6)}$$

We show that both the variance and covariance terms appearing in (A.6) are dominated by $c(\tilde{I}(k)+\tilde{I}(j))$.

Given $\mathscr{G}^{(n)}$, $N_{f,s,i}(\tau_k)$ is, for large n, approximately $\operatorname{bin}(m_f,\, p_k(s,\, i))$ so $\operatorname{var}(N_{f,s,i}(\tau_k)|\mathscr{G}^{(n)}) \approx m_f p_k(s,\, i)(1-p_k(s,\, i))$. By the last inequality in (A.5) this implies $\operatorname{var}(N_{f,s,i}(\tau_k)|\mathscr{G}^{(n)}) \leqslant c\tilde{I}(k)$.

In the second sum of (A.6), $\tau_j<\tau_k$, so if we condition on $\mathscr{F}^{(n)}_{\tau_j}$ we have

$$E(N_{f,s,i}(\tau_j)N_{f,s,i}(\tau_k)|\mathscr{G}^{(n)}) = E(N_{f,s,i}(\tau_j)E(N_{f,s,i}(\tau_k)|\mathscr{F}^{(n)}_{\tau_j},\, \mathscr{G}^{(n)})|\mathscr{G}^{(n)}),$$

and

$$E(N_{f,s,i}(\tau_k)|\mathscr{F}^{(n)}_{\tau_j},\, \mathscr{G}^{(n)}) = \sum_{s',i'} p_{j,k}(s',\, i';\, s,\, i)N_{f,s',i'}(\tau_j),$$

where $p_{j,k}(s',\, i';\, s,\, i)$ is the $\mathscr{G}^{(n)}$-conditional probability that a size f family in state $(s',\, i')$ at τ_j will be in $(s,\, i)$ at τ_k (we will not need to derive what this probability equals). Using this, one can show that

$$\operatorname{cov}(N_{f,s,i}(\tau_j),\, N_{f,s,i}(\tau_k)|\mathscr{G}^{(n)}) =$$
$$m_{f_{(2)}}\binom{f}{i\quad s}\sum_{s',i'}\frac{(\tilde{S}(j)-s')_{(s)}(\tilde{I}(j)-i')_{(i)}(\tilde{R}(j)-r')_{(r)}}{(\mu n-f)_{(f)}}\,p_j(s',\, i')p_{j,k}(s',\, i';\, s,\, i)$$
$$+\, m_f p_j(s,\, i)p_{j,k}(s,\, i;\, s,\, i) - m_f^2 p_j(s,\, i)p_k(s,\, i). \qquad \text{(A.7)}$$

The two terms with m_f to the first power (remember that $m_{f_{(2)}} = m_f^2 - m_f$) are both less than $c\tilde{I}(j)$. It remains to show that the absolute value of the difference of the two m_f^2-terms is bounded by $c\tilde{I}(k)$. But $\sum_{s',i'} p_j(s',\, i')p_{j,k}(s',\, i';\, s,\, i) = p_k(s,\, i)$,

$$\frac{(\tilde{S}(j)-s')_{(s)}(\tilde{I}(j)-i')_{(i)}(\tilde{R}(j)-r')_{(r)}}{(\mu n-f)_{(f)}} = \frac{\tilde{S}(j)_{(s)}\tilde{I}(j)_{(i)}\tilde{R}(j)_{(r)}}{(\mu n)_{(f)}}\left(1+O\left(\frac{1}{\tilde{I}(j)}\right)\right)$$

and

$$\binom{f}{i\quad s}\tilde{S}(j)_{(s)}\tilde{I}(j)_{(i)}\tilde{R}(j)_{(r)}/(\mu n)_{(f)} = p_j(s,\, i).$$

By (A.5) it hence follows that the value of the difference is less than $m_f^2(c/\tilde{I}(j))(\tilde{I}(j)/\mu n)^i(\tilde{I}(k)/\mu n)^i \leqslant c\tilde{I}(k)$.

Thus, by (A.6)

$$\operatorname{var}(X^{(n)}|\mathscr{G}^{(n)}) \leqslant \frac{c}{(\log n)^2}\sum_{k=0}^{\kappa}\left(\frac{\tau_{k+1}-\tau_k}{\tilde{I}(k)}\right)^2\tilde{I}(k)$$
$$+\frac{c}{(\log n)^2}\sum_{0\leqslant j<k<\kappa}\frac{(\tau_{k+1}-\tau_k)(\tau_{j+1}-\tau_j)}{\tilde{I}(j)\tilde{I}(k)}(\tilde{I}(j)+\tilde{I}(k))$$
$$\leqslant \frac{ct\log n}{(\log n)^2}\int_0^{t\log n}\frac{1_{\{1(u)>0\}}}{I(u)}\,du = ct\int_0^t\frac{1_{\{I(u\log n)>0\}}}{I(u\log n)}\,du.$$

By lemma 4 the integral converges to 0 in probability.

Proof of lemma 3. Similar to $\langle M_1^{(n)}\rangle$ and $\langle M_2^{(n)}\rangle$ we have

$$\langle M_3^{(n)}\rangle(t)=\frac{\lambda\mu N}{N-1}\sum_{f,s,i}\frac{i^2 s}{(f-1)^2}\int_0^t 1_{A_3^{(n)}(u\log n)}\frac{N_{f,s,i}^{(n)}(u\log n)}{I^{(n)}(u\log n)}\,du.$$

On each integral term above with $i\geqslant 1$ we can apply lemma 4 (with $\beta=1/2$ and arbitrary small ϵ) which together with lemma 5 implies that the integral term converges in probability to

$$\left(\begin{array}{c} f \\ i\quad s\end{array}\right)\frac{p_f}{\mu}\times\begin{cases} 0 & \text{if } i>1\\ ((t-t_1)^+\wedge t_1)+(1-\pi)^{f-1}((t-2t_1)^+\wedge t_2) & \text{if } i=1,\ s=f-1\\ (1-\pi)^s\pi^{f-1-s}((t-2t_1)^+\wedge t_2) & \text{if } i=1,\ s<f-1.\end{cases}$$

Finally, if we multiply these terms by $\lambda\mu i^2 s/(f-1)^2$, their sum equals

$$\alpha((t-t_1)^+\wedge t_1)+(1-\pi)\alpha((t-2t_1)^+\wedge t_2),$$

where $\alpha=\lambda\sum_{f>1}fp_f/(f-1)$. This is what was claimed in the lemma.

Proof of Corollary. By th. VI.1.1. and VI.1.2. in Andersen *et al.* (1993) it follows that

$$\hat\lambda_t^{(n)}\overset{P_0^{(n)}}{\to}\lambda,\ \tilde\gamma_t^{(n)}\overset{P_0^{(n)}}{\to}\gamma\quad\text{and}\quad\sqrt{n}(\hat\lambda_t^{(n)}-\lambda)\overset{D}{\to}Z,\text{ as } n\to\infty,$$

where Z is normally distributed with mean 0 and variance $\pi\mu$. From this the consistency statement follows and also that $\sqrt{n}(\hat\lambda_t^{(n)}-\lambda)$ is bounded in probability. We know that $M^{(n)}\Rightarrow M$, so the statement $\hat M^{(n)}\Rightarrow M$ will follow by Slutsky's theorem if we can show that

$$\hat M^{(n)}(t)-M^{(n)}(t)\overset{P_0^{(n)}}{\to}0$$

for each t, and since

$$\hat\lambda_t^{(n)}\overset{P_0^{(n)}}{\to}\lambda$$

this will follow if

$$\hat M^{(n)}(t)/\hat\lambda_t^{(n)}-M^{(n)}(t)/\lambda\overset{P_0^{(n)}}{\to}0$$

which we now show. From the definition of $M^{(n)}$ and $\hat M^{(n)}$,

$$\frac{\hat M^{(n)}(t)}{\hat\lambda_t^{(n)}}-\frac{M^{(n)}(t)}{\lambda}=\sqrt{n}\left(\frac{1}{\hat\lambda_t^{(n)}}-\frac{1}{\lambda}\right)\frac{\mu}{\sqrt{\log n}}\sum_{i:f_i>1}\int_0^{t\log n}\frac{I_i^{(n)}(s-)}{(f_i-1)I^{(n)}(s-)}\,dN_i^{(n)}(s).$$

The factor $\sqrt{n}(1/\hat\lambda_t^{(n)}-1/\lambda)$ is bounded in probability, so this converges to 0 in probability if

$$(\log n)^{-1/2}U_n\overset{P_0^{(n)}}{\to}0,$$

where

$$U_n:=\sum_{i:f_i>1}\int_0^\infty\frac{I_i^{(n)}(s-)}{(f_i-1)I^{(n)}(s-)}\,dN_i^{(n)}(s).$$

Let τ_k be as above and $E_k:=\{\tilde I(k)-\tilde I(k-1)=1\}$ is the event that the kth jump is an infection. On E_k we let $\chi_k=i$ if $\tilde I_i(k)-\tilde I_i(k-1)=1$, $i=1,\ldots,n$. This means that χ_k is a marker for which family received the infection. We have assumed there is one infective at the start of the epidemic so if there ultimately are $N(\infty)$ infected there will in all be $2N(\infty)-1$ jumps. An alternative way to write the sum of integrals above is thus

$$U_n := \sum_{i:f_i>1} \int_0^\infty \frac{I_i^{(n)}(s-)}{(f_i-1)I^{(n)}(s-)}\, dN_i^{(n)}(s) = \sum_{k=1}^{2N(\infty)-1} 1_{E_k} \frac{\tilde{I}_{\chi k}(k-1)}{(f_{\chi k}-1)\tilde{I}(k-1)}. \tag{A.8}$$

Further

$$E\left(1_{E_k} \frac{\tilde{I}_{\chi k}(k-1)}{f_{\chi k}-1} \middle| \mathscr{F}_{\tau_{k-1}}, \mathscr{G}\right) = 1_{E_k} \sum_{i=1}^{n} \frac{\tilde{I}_i(k-1)}{f_i-1} \frac{\tilde{S}_i(k-1)}{\tilde{S}(k-1)} \leqslant \frac{\tilde{I}(k-1)}{\tilde{S}(k-1)}. \tag{A.9}$$

Remember that $\mathscr{F}_t$ is the σ-algebra generated by the epidemic process up to t if we observe the population on *family* level and $\mathscr{G}$ is the σ-algebra generated by the *whole* epidemic process but now only observing the *population totals*.

By (A.8) and (A.9), $E(U_n) \leqslant \sum_k (1_{E_k}/\tilde{S}(k-1)) \leqslant N(\infty)/S(\infty)$. As mentioned in section 3, the proportion infected ($= N(\infty)/\mu n = 1 - S(\infty)/\mu n$) is concentrated around π in case of a major epidemic. So if we let $B_n = \{S^{(n)}(\infty)/\mu n \geqslant (1-\pi)/2\}$ it follows that $P(B_n) \to 1$. Pick $\epsilon > 0$ arbitrary. Then

$$\begin{aligned} P\left(\frac{1}{\sqrt{\log n}} U_n > \epsilon\right) &= P(1_{B_n} U_n > \epsilon\sqrt{\log n}) + P(1_{B_n^C} U_n > \epsilon\sqrt{\log n}) \\ &\leqslant \frac{E(1_{B_n} U_n)}{\epsilon\sqrt{\log n}} + P(B_n^C) \leqslant \frac{1+\pi/2}{(1-\pi/2)\epsilon\sqrt{\log n}} + P(B_n^C) \to 0. \end{aligned}$$