

# Who was the infector—probabilities in the presence of variability in latent and infectious times

Åke Svensson

Received: 30 August 2012 / Revised: 30 January 2013 / Published online: 26 February 2013  
© Springer-Verlag Berlin Heidelberg 2013

**Abstract** The probability that an observed infection has been transmitted from a particular member of a set of potential infectors is calculated. The calculations only use knowledge of the times of infection. It is shown that the probabilities depend on individual variability in latent and infectious times. The analysis are based on different background information and different assumptions on the progress of infectivity. The results are illustrated by numerical calculations and simulations.

**Keywords** Epidemic models · Generation times · Transmission chains

**Mathematics Subject Classification** 62M99 · 62P10

## 1 Introduction

When analyzing the spread of an infectious disease it is often of interest to identify the infector of an infected person. This can seldom be done with complete certainty. However, it may be possible to calculate the probability that a certain member of a set of potential infectors is the true transmitter. How this can be done, using only observations of the times of infection, is the topic of the present paper.

As in all statistical and probabilistic analysis we have to carefully consider how the observed data are generated. In Sect. 3 we assume that the histories of the potential infectors are unrelated, and in Sect. 4 that the candidates form a transmission tree, i.e., the set of possible infectors consists of one original infected and a sequence of persons that have infected each other.

One reason to calculate probabilities is to better understand the transmission dynamics. In a study of SARS outbreaks [Wallinga and Teunis \(2004\)](#) analysed the possible

---

Å. Svensson (✉)

Department of Mathematics, Stockholm University, Stockholm, Sweden  
e-mail: akes@math.su.se

impact of control measures. Given times of infection, they could by adding the probabilities that a certain infector was responsible for future infections, estimate the expected number of secondary infections. [Cauchemez and Ferguson \(2012\)](#) also study the related, and more complex, problem of how to find the most probable transmission chain. In their analysis variations in infectivity of potential infectors were related to observable quantities.

Understanding transmission routes has become an important tool in the analysis of epidemic outbreaks. E.g. [Cauchemez et al. \(2011\)](#) study how transmission of influenza is affected by social networks based on data from a community outbreak, and [Hens et al. \(2012\)](#) analyse a school-based outbreak using auxiliary information on possible transmission routes. These analysis are also based on models with non-individual infectivity.

The probabilities, that we are interested in will depend on generation times, i.e. the times between a primary infection and its corresponding secondary infections. Generation times and the related concept of contact intervals are discussed by e.g. [Svensson \(2007\)](#), [Tomba et al. \(2010\)](#) and [Kenah \(2011\)](#). Individual random variations in infectivity will have substantial impact. In the examples used in this paper the variations are assumed to be generated by a SEIR model with random latent and infectious periods. Basic assumptions also concern homogeneous mixing, and constant infectivity during the infectious period. Of course, it is possible to analyze more complicated models, but our main purpose here is to illustrate potential consequences of individual variations and to suggest possibilities to perform calculations. For this reason we use a simple setting. The models and the notation are presented in Sect. 2.

Section 5 gives numerical examples that illustrate that the assumptions underlying the analysis are crucial. In the simulated examples, as will certainly be the case for observations from real epidemics, the probabilities found, do in general not give a very precise indication of who was the infector for a specific case. Thus other information than the times of infection is required to precisely indicate the infector.

In Sect. 6 how the findings in the numerical examples may be generalized is discussed.

## 2 Basic model and notation

We will study a situation where  $v$  infections are observed to occur at times  $\tau_1 < \tau_2 < \dots < \tau_v$ . Without loss of generality we may assume that  $\tau_1 = 0$ .

### 2.1 Infectivity and generation times

It is assumed that an infected person spreads the infection according to an intensity process that depends on the time after the infection, i.e. the age of the infection. The intensity processes may be individual and random, but the random intensity functions for different individuals are assumed to be independent. We consider spread in a closed population that is assumed to be homogeneously mixing. The victim for a new infection is a randomly chosen member of the population.

The intensity process for the  $i$ 'th infected is denoted by  $\kappa_i$ . The interpretation is that the infected person has potentially infectious contacts according to a Poisson process with time-varying intensity  $\kappa_i(a)$ , where  $a$  is the age of the infection. The contact will lead to a transmission if the contacted is susceptible. If there exists immunity in the population the occurrence of new infections will be influenced by this. Let  $s(t)$  be the proportion of susceptible persons in the population at time  $t$ . If the  $i$ 'th infected is infected at time  $\tau_i$  this person will cause secondary infections according to a Poisson process with intensity  $s(t)\kappa_i(t - \tau_i)$ ,  $t > \tau_i$ .

In this model the total infectivity that individual  $i$  spreads is

$$\lambda_i = \int_0^\infty \kappa_i(a) da \tag{1}$$

The parameter  $\lambda_i$  can be interpreted as the mean number of possible infectious contacts this infected individual makes.

The basic reproduction number, i.e. the expected number of secondary infections in a totally susceptible population, is then

$$R_0 = E(\lambda_i) = \int_0^\infty g(a) da, \tag{2}$$

where

$$g(a) = E(\kappa_i(a)). \tag{3}$$

The cohort (or basic) generation time density,  $k$  (cf [Svensson 2007](#); [Tomba et al. 2010](#)), corresponds to  $g$  normalized to have total mass 1, i.e.,

$$k(a) = \frac{E(\kappa_i(a))}{R_0} = \frac{g(a)}{R_0}. \tag{4}$$

The mean generation time is defined as

$$\int_0^\infty ak(a) da. \tag{5}$$

### 2.2 Models for individual variations

In order to illustrate the effects of individual variability we will use a model of the type generally referred to as a SEIR model. An infection is assumed to be followed by a latent period, during which the infection is not transmitted. After the latent period follows an infectious period. We will here, for simplicity, assume that a person makes

infectious contacts with a constant non-random rate  $\beta$ , that is the same for all infected. If an infectious person makes such a contact with a susceptible person the infection is transmitted.

The duration of the latent and infectious periods may vary between individuals. Let  $L_i$  be the duration of the latent period and  $X_i$  the duration of the infectious period of the  $i$ 'th infected. Furthermore let  $I_i(a)$  be the indicator function that the  $i$ 'th infected is infectious at time  $a$  after the infection, i.e.,

$$I_i(a) = \begin{cases} 1 & \text{if } L_i \leq a \leq L_i + X_i, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\kappa_i(a) = \beta I_i(a). \tag{6}$$

Obviously  $\lambda_i = \beta X_i$  and  $R_0 = \beta E(X)$ .

Assume that the pairs  $L_i$  and  $X_i$  are independent and that their respective distribution functions are  $H_L$  and  $H_I$ , with densities  $h_L$  and  $h_I$ , then

$$k(a) = \frac{H_L(a) - \int_0^a H_L(a-s)h_I(s)ds}{E(X)}. \tag{7}$$

If there is no latent period, i.e. when the infectious period starts immediately after infection, and

$$k(a) = \frac{1 - H_I(a)}{E(X)}. \tag{8}$$

Furthermore

$$p(a) = \Pr(I_i(a) = 1) = E(X)k(a), \tag{9}$$

and the probability that the  $i$ 'th infected is infectious at time  $t$  is

$$p(t - \tau_i) = E(X)k(t - \tau_i).$$

As examples we will use three different cases where the latent and infectious periods are independent and gamma-distributed. To make the cases comparable we will choose parameter values so that the mean length of the latent and the mean generation times in each case are approximately what is assumed for seasonal influenza, with day as the time unit (cf Carrat et al. 2008).

In the first case both the latent and the infectious periods are assumed to be exponentially distributed. This implies a large individual variation. In the second case the variation is smaller, and in the third case the times are assumed to be constant. Thus in the third case there is no individual variation.

**Case 1:**  $L_i$  is exponential distributed with intensity  $\mu_L$ , and  $X_i$  is exponential distributed with intensity  $\mu_I$ . If  $\mu_L \neq \mu_I$

$$k(a) = \frac{\mu_L \mu_I}{\mu_I - \mu_L} (\exp(-\mu_L a) - \exp(-\mu_I a)),$$

and

$$p(a) = \frac{\mu_L}{\mu_I - \mu_L} (\exp(-\mu_L a) - \exp(-\mu_I a)).$$

If  $\mu = \mu_L = \mu_I$  then

$$k(a) = \mu^2 a \exp(-\mu a),$$

and

$$p(a) = \mu a \exp(-\mu a).$$

The mean generation time is  $1/\mu_L + 1/\mu_I$ . In the calculations we have chosen  $\mu_L = 1$ , and  $\mu_I = 1/2$ . This gives the mean generation time 3.

**Case 2:**  $L_i$  is gamma distributed with shape parameter  $\alpha$  and rate parameter  $\mu_L \alpha$ , and  $X_i$  is also gamma distributed with shape parameter  $\delta$  and rate parameter  $\mu_I \delta$ . The means of the latent and infectious periods are  $1/\mu_L$  and  $1/\mu_I$ . The mean generation time is  $1/\mu_L + 1/\mu_I \frac{1+\delta}{2\delta}$  (cf Svensson 2007). In the calculations we have chosen  $\alpha = \delta = 8$ ,  $\mu_L = 1$ , and  $\mu_I = 9/32$ . This gives the mean generation time 3.

**Case 3:**  $L_i$  and  $X_i$  are constant. This corresponds to  $\alpha = \delta = \infty$ . To obtain the same mean latent and generation times as in the two previous cases we chose  $L_i \equiv 1$  and  $X_i \equiv 4$ .

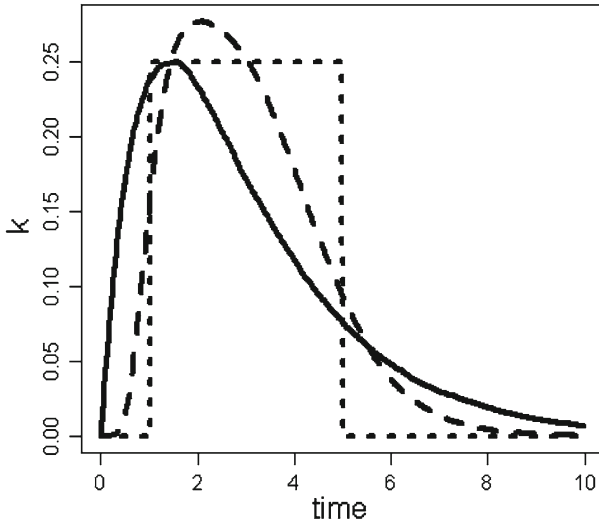
Figure 1 illustrates the cohort generation time density for the three cases. It can be observed that large variability results in long tails of the generation time density.

### 3 Unrelated potential infectors

In this section we will assume that the first  $v - 1$  infections are unrelated and derive the probability that the  $i$ 'th infected infects the  $v$ 'th. Let  $t = \tau_v$ .

#### 3.1 Non-random infectivity

Suppose that  $v - 1$  Poisson process are running in parallel and that they have intensity functions  $v_1(t), \dots, v_{v-1}(t)$ . Given that an event happens in one of these processes at time  $t$  the probability that it occurs in the  $i$ 'th process is



**Fig. 1**  $k(a)$ , for case 1 (solid), case 2 (dashed), and case 3 (dotted)

$$\frac{v_i(t)}{\sum_j v_j(t)}$$

(The summation is for  $j = 1, \dots, v - 1$ ).

In case when all individuals have the same infectivity intensity the probability that the  $i$ 't infector is responsible for the  $v$ 'th infection, at time  $t = \tau_v$ , can be calculated as

$$P_i^a = \frac{s(t)k(t - \tau_i)}{\sum_j s(t)k(t - \tau_j)} \tag{10}$$

or equivalently (according to (9))

$$P_i^a = \frac{p(t - \tau_i)}{\sum_j p(t - \tau_j)} = wp(t - \tau_i), \tag{11}$$

where  $w$  is a constant such that  $\sum P_i^a = 1$ .

As will be clear from the following discussion the expression is not valid if the intensity processes are random. In such cases the probabilities (10) and (11) can, at best, be regarded as approximations.

### 3.2 Random infectivity

We will now consider the possibility that the infectious processes are random. Due to the assumption of homogeneous mixing and the assumption that all infectious individuals are equally infectious during their infectious periods

$$Z_i = \frac{I_i(t - \tau_i)}{\sum_j I_j(t - \tau_j)} \tag{12}$$

is, conditional on the latent and infectious times, the probability that the infection is transmitted from infector  $i$ . Without the conditioning  $Z_i$  should be regarded as a random variable.

Since there has to be at least one infector it is a necessary restriction that

$$\sum_j I_j(t - \tau_j) > 0.$$

If there is no other relation between the potential infectors this is the only restriction that has to be considered.

The probability that the  $i$ 'th infected is the infector can be expressed as:

$$\begin{aligned} P_i^u &= E \left( Z_i \mid \sum_j I_j(t - \tau_j) > 0 \right) \\ &= \frac{\sum_{r=1}^{v-1} \frac{1}{r} \Pr \left( I_i(t - \tau_i) = 1, \sum_{j \neq i} I_j(t - \tau_j) = r - 1 \right)}{\Pr \left( \sum_j I_j(t - \tau_j) > 0 \right)} \\ &= p(t - \tau_i) \frac{E \left( \frac{1}{1 + \sum_{j \neq i} I_j(t - \tau_j)} \right)}{\Pr \left( \sum_j I_j(t - \tau_j) > 0 \right)} \\ &= p(t - \tau_i) w_i. \end{aligned} \tag{13}$$

Due to construction (cf Eq. 12)  $\sum P_i^u = \sum p(t - \tau_i) w_i = 1$ . Comparing the probabilities  $P_i^u$  and  $P_i^a$  we find that

$$\frac{P_i^u}{P_i^a} \sim w_i. \tag{14}$$

Since the factors  $w_i$  depend on  $i$  the probabilities will differ from those given by (11). Note that  $\sum_{j \neq i} I_j(t - \tau_j)$  is the sum of  $v - 2$  independent random variables. These random variables are stochastically ordered according to the probabilities  $p(t - \tau_j)$ , which are proportional to  $k(t - \tau_j)$ . Thus, the sum tends to be large when  $k(t - \tau_i)$  is small.

If there is no latent time the ordering of the  $\tau_i$ 's imply that  $k(t - \tau_i)$  and thus  $w_i$  increases with  $i$  (cf 8). As a consequence the probabilities  $P_i^a$  are, in this case, too large for long generation times.

### 4 Spread in a transmission tree

A possible scenario is that we know that the infected and the transmission links form a tree with its root at the first infected. This can be the case if the observations come from a study of infectious spread within a family, a school class, or some other small closed population. If we observe a transmission tree we know that at least one of the candidates is infectious at any time of infection  $\tau_2, \dots, \tau_\nu$ . This means that

$$\sum_{j < m} I_j(\tau_m - \tau_j) > 0.$$

for all  $2 \leq m \leq \nu$ .

A transmission tree may be a part of a larger transmission chain. In order to have a well defined situation we will assume that the observed infections are the first emerging from an initial infector. We will then have another restriction, namely that the potential infectors represent all infections before time  $t$ . If this is the case we can also calculate the susceptible proportion of the population. If the population has  $n$  susceptible members at the time of the initial infection the assumption that contacts are made uniformly at random in the population leads to that  $s(t) = 1 - (i - 1)/n$  when  $\tau_i \leq t < \tau_{i+1}$ .

In this setting it is of interest to calculate the probability that the  $i$ 'th infected infected the  $j$ 'th. Let

$$Z_{ij} = \frac{I_i(\tau_j - \tau_i)}{\sum_m I_m(\tau_j - \tau_m)}. \tag{15}$$

The probabilities are

$$P_{ij}^c = E(Z_{ij} \mid \sum_{j < m} I_j(\tau_m - \tau_j) > 0 \text{ if } m \leq \nu \text{ and only } \nu \text{ infected up to } \tau_\nu). \tag{16}$$

We have not been able to find any simple closed version. In the following section we suggest a simulation procedure to do the calculations. It turns out that the probabilities (16) will depend on  $R_0$  (via  $\beta$ , i.e. the rate of potential infectious contacts). Since the actual infectivity in a population also depends on the proportion,  $s(t)$ , of susceptible individuals the probabilities will also depend on the population size.

#### 4.1 A simulation procedure

We first assume that the infectivity functions,  $(\kappa_1, \dots, \kappa_{\nu-1})$  for the first  $\nu - 1$  infected in the chain are known. The total infectious force at time  $\tau \leq \tau_\nu$  is

$$U(\tau) = \beta s(\tau) \sum_{i=1}^{\nu-1} \kappa_i(\tau - \tau_i).$$



We start by deriving a density for  $\xi = (\tau_2, \dots, \tau_\nu)$ . Infections occur according to a Poisson process with the random intensity process  $U$ . It is random because it depends dynamically on when previous infections occurred. In fact, if we subtract the integrated infectious force from the process that counts the number of infections from the initiation and forwards we obtain a martingale. From general theory, (cf [Bremaud 1981](#), pg 226) it follows that the density (related to a standard Poisson process) of the random vector  $\xi$  is

$$\prod_{i=2}^{\nu} U(\tau_i) \exp \left( \int_0^{\tau_\nu} (1 - U(\tau)) d\tau \right).$$

The probability that it is the  $i$ 'th infector that causes the  $j$ 'th infection is

$$\frac{\beta s(\tau_j) \kappa_i(\tau_j - \tau_i)}{U(\tau_j)} = \frac{\kappa_i(\tau_j - \tau_i)}{\sum_m \kappa_m(\tau_j - \tau_m)}.$$

If we take into consideration that the infectivity functions are random we find that the conditional expectations can be expressed as

$$P_{ij}^c = \frac{E \left( \frac{\beta s(\tau_j) \kappa_i(\tau_j - \tau_i)}{U(\tau_j)} \prod_{k=2}^{\nu} U(\tau_k) \exp(-\int_0^{\tau_\nu} U(\tau) d\tau) \right)}{E \left( \prod_{i=2}^{\nu} U(\tau_i) \exp(-\int_0^{\tau_\nu} U(\tau) d\tau) \right)}.$$

Let  $(\kappa_1^r, \dots, \kappa_{\nu-1}^r)$ ,  $r = 1, \dots, m$  be  $m$  sets of simulated infectivity functions, and let  $U^r$  be the corresponding total infectivity functions. We can estimate  $P_{ij}^c$  with

$$\hat{P}_{ij}^c = \frac{\sum_{r=1}^m \frac{\beta s(\tau_j) \kappa_i^r(\tau_j - \tau_i)}{U^r(\tau_j)} \prod_{k=2}^{\nu} U^r(\tau_k) \exp(-\int_0^{\tau_\nu} U^r(\tau) d\tau)}{\sum_{r=1}^m \prod_{i=2}^{\nu} U^r(\tau_i) \exp(-\int_0^{\tau_\nu} U^r(\tau) d\tau)}.$$

Observe that if some  $U^r(\tau_j) = 0$ ,  $j = 2, \dots, \nu$ , then there cannot (with probability 1) exists a chain with the given times of infection. Simulated values of the infectivity force function that leads to this will not give any contribution to the estimate.

For the models described in Sect. 2 it is enough to know the latent and infectious times to find the infectivity functions. Let  $W = (L_1, X_1, \dots, L_{\nu-1}, X_{\nu-1})$ . Furthermore let  $J_i(\tau, W)$  equal 1 if the  $i$ 'th infected is infectious at time  $\tau$ , which it is when  $\tau_i + L_i \leq \tau < \tau_i + L_i + X_i$ , and 0 otherwise. The number of the potential infectors that are infectious at time  $\tau$  is  $J(\tau, W) = \sum J_i(\tau, W)$ , and  $U(\tau) = \beta s(\tau) J(\tau, W)$ .

Let  $a_{i,j}$  be the total time the  $i$ 'th infector is infectious before the  $j$ 'th infection, then

$$a_{i,j} = \max(\min(\tau_j - L_i - \tau_i, X_i), 0).$$

Of course  $a_{i,j}$  can only be positive if  $i < j$ . Since  $s(\tau_j) = 1 - (j - 2)/n$  it follows that the sum of all infectious times up till time  $\tau_v$  can be calculated as

$$T(W) = \sum_{i=1}^{v-1} \sum_{j=1}^{v-1} \left(1 - \frac{j-1}{n}\right) (a_{i,j+1} - a_{i,j}).$$

Now let  $W^1, \dots, W^m$  be a sequence of independent simulated random elements reflecting  $v - 1$  latent and infectious times. The probability that the  $i$ 'th infector is the one who infects the  $j$ 'th at time  $\tau_j$  can be estimated by

$$\frac{\sum_{r=1}^m \frac{J_i(\tau_j, W^r)}{J(\tau_j, W^r)} \prod_{k=2}^v J(\tau_k, W^r) \exp(-\beta T(W^r))}{\sum_{r=1}^m \prod_{k=2}^v J(\tau_k, W^r) \exp(-\beta T(W^r))}.$$

In the following calculations we have chosen to make sufficiently many simulation so that

$$\sum_{i=1}^m \mathbb{I}\left(\prod_{j=2}^v J(\tau_j) > 0\right) = k_m,$$

where  $k_m$  is a predesigned number. Thus the simulations have to produce  $k_m$  possible chains.

### 5 Numerical examples

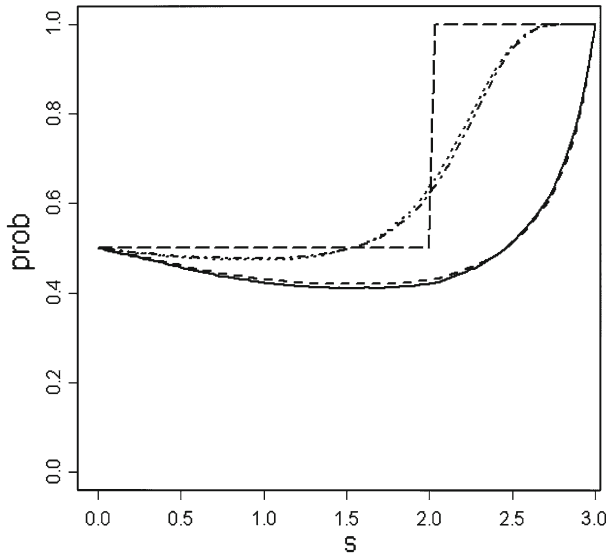
The expressions  $P_i^a$ ,  $P_i^u$  and  $P_{ij}^c$  depend on  $v$  and  $(\tau_1, \dots, \tau_v)$  as well as on the assumed cohort generation time.  $P_{ij}^c$  also depends on  $\beta$  or equivalently on  $R_0$ .

To illustrate this we first consider a situation where there are only three infected. The second example illustrates a more complicated situation with a longer observed transmission tree.

#### 5.1 Two possible infectors

The probability that the first infected is the true infector will depend both on  $\tau_3 = t$ , i.e., the time when the third infected was infected and  $\tau_2 = s$ , the time for the second infection.

The calculations are relatively simple if there is no latency time and the infectious periods are exponentially distributed with rate  $\mu_I$ . In this case explicit expression of the probabilities that the initial infected infects the third are



**Fig. 2** Probability that the first infected has infected the third infected as a function of the time of infection for the second infected when the third infection occurs at time  $t = 3$ .  $P_1^a$  case 1 (solid) and case 2 (dashed).  $P_1^u$  case 1 (dotted) and case 2 (dotdashed). For case 3,  $P_1^a = P_1^u$  longdash

$$P_1^a = \frac{1}{1 + \exp(\mu_I s)},$$

$$P_1^u = \frac{1 - \exp(-\mu_I(t - s))/2}{1 + \exp(\mu_I s) - \exp(-\mu_I(t - s))},$$

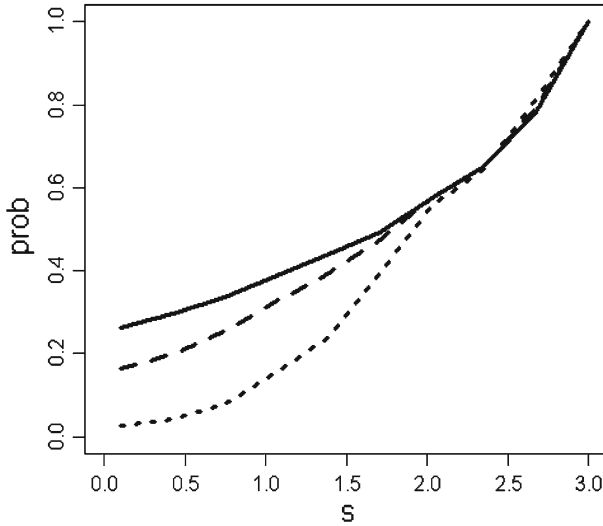
and

$$P_{13}^c = 1/2.$$

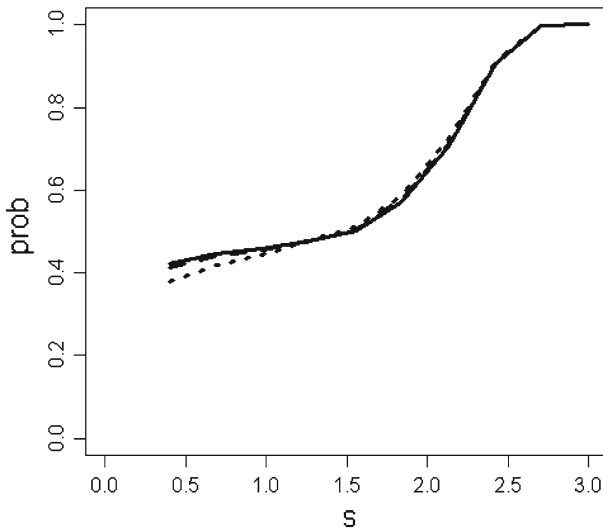
These three probabilities are concerned with the same event but calculated under different assumptions. Observe that they are all different and ordered as  $P_1^u \leq P_1^a \leq P_{13}^c$ .

If there is a latent period the calculations are more complicated. We will consider the three cases defined in Sect. 2. We assume that the population is large, so that the fact that one person is infected does not influence the probability of further infections. The probabilities,  $P_1^a$  and  $P_1^u$  are illustrated in Fig. 2. The difference between  $P_1^a$  and  $P_1^u$  is relatively small but depends heavily on the generation time model used. If there is no individual variability  $P_1^a = P_1^u$ .

Figures 3 and 4 illustrate the probabilities in case the infections form a tree. It is seen that the influence of  $R_0$  is considerable in cases where there is a substantial individual variability. As can be expected it is less in case 2 than in case 1. In case 2 the relation is illustrated only for  $\tau_2 > 0.4$ . The probability for smaller time distance between the first and second infection will be small. If there is no individual variability, as in case 3,  $P_1^a = P_1^u = P_{13}^c$ .



**Fig. 3** Probability that the first infected has infected the third infected as a function of the time of infection for the second infected ( $s > 0.4$ ) when the third infection occurs at time  $t = 3$ .  $P_{13}^c$  for  $R_0 = 1.2$  (solid), 2 (dashed) and 4 (dotted). Model as in case 1



**Fig. 4** Probability that the first infected has infected the third infected as a function of the time of infection for the second infected ( $s > 0.4$ ) when the third infection occurs at time  $t = 3$ .  $P_{13}^c$  for  $R_0 = 1.2$  (solid), 2 (dashed) and 4 (dotted). Model as in case 2

5.2 Observation of a transmission tree

To illustrate the use of the calculations in a more complicated situation we will use a simulated example. The simulated epidemic takes place in a large population.

This means that we do not have to consider effects of changed number of susceptible members in the population. The spread is simulated with a model as in case 2 described above and with  $R_0 = 2$ . The times of infection are

$$(\tau_1, \dots, \tau_{10}, \tau_{11}) = (0, 0.751, 1.622, 3.009, 3.248, 3.290, 3, 527, 3.963, 4.158, 4.371, 4.425).$$

In sequence the infectors were (the first infection comes from outside the population)

$$(-, 1, 2, 3, 3, 3, 4, 7, 4, 8, 4).$$

This means that the eleventh infection occurred at  $t = 4.425$  and that the real infector in this simulated example was the fourth infected.

We can now calculate the probabilities that the  $i$ 'th infected infected the  $j$ 'th infected ( $i < j$ ) given that the eleven observed infections form a tree. The probabilities for case 1 are presented in Table 1, for case 2 in Table 2, and for case 3 in Table 3. The values of  $P_{ij}^c$  are derived from simulations as described above with  $k_m = 5,000$  and with  $R_0 = 2$ . In case 3, with no individual variation, there is no need for simulations, since all possible infectors are equally likely. This also implies that  $R_0$  does not influence the probabilities.

In order to illustrate the importance of  $R_0$  for the probabilities Table 4 gives the estimates of the expected number of infections due to the different infected under the different models and for different values of  $R_0$ . These numbers are the sums of the probabilities related to the  $i$ 'th infected. The results differ for the three cases. If there is much variability, as in case 1, there is a considerable effect of  $R_0$ .

**Table 1** Estimated probabilities that the  $i$ 'th infected infected the  $j$ 'th infected, using the model as in case 1 with  $R_0 = 2$

i	j									
	2	3	4	5	6	7	8	9	10	11
1	1	0.66	0.27	0.22	0.21	0.17	0.11	0.09	0.08	0.07
2	-	0.34	0.36	0.30	0.29	0.23	0.16	0.13	0.11	0.11
3	-	-	0.37	0.34	0.33	0.28	0.20	0.17	0.14	0.13
4	-	-	-	0.14	0.15	0.17	0.17	0.16	0.15	0.14
5	-	-	-	-	0.02	0.08	0.13	0.14	0.14	0.13
6	-	-	-	-	-	0.07	0.13	0.14	0.14	0.13
7	-	-	-	-	-	-	0.10	0.12	0.12	0.12
8	-	-	-	-	-	-	-	0.05	0.08	0.09
9	-	-	-	-	-	-	-	-	0.05	0.06
10	-	-	-	-	-	-	-	-	-	0.01

**Table 2** Estimated probabilities that the  $i$ 'th infected infected the  $j$ 'th infected, using the model as in case 2 with  $R_0 = 2$

i	j									
	2	3	4	5	6	7	8	9	10	11
1	1	0.77	0.33	0.31	0.31	0.28	0.19	0.14	0.11	0.10
2	–	0.23	0.36	0.35	0.35	0.34	0.26	0.21	0.16	0.15
3	–	–	0.31	0.34	0.35	0.35	0.28	0.23	0.19	0.18
4	–	–	–	0.00	0.00	0.02	0.14	0.17	0.18	0.18
5	–	–	–	–	0.00	0.00	0.07	0.11	0.15	0.15
6	–	–	–	–	–	0.00	0.05	0.10	0.14	0.14
7	–	–	–	–	–	–	0.01	0.04	0.08	0.09
8	–	–	–	–	–	–	–	0.00	0.00	0.01
9	–	–	–	–	–	–	–	–	0.00	0.00
10	–	–	–	–	–	–	–	–	–	0.00

**Table 3** Estimated probabilities that the  $i$ 'th infected infected the  $j$ 'th infected, using the model as in case 3

i	j									
	2	3	4	5	6	7	8	9	10	11
1	1	1	1/3	1/3	1/3	1/3	1/3	1/4	1/6	1/6
2	–	0	1/3	1/3	1/3	1/3	1/3	1/4	1/6	1/6
3	–	–	1/3	1/3	1/3	1/3	1/3	1/4	1/6	1/6
4	–	–	–	0	0	0	0	1/4	1/6	1/6
5	–	–	–	–	0	0	0	0	1/6	1/6
6	–	–	–	–	–	0	0	0	1/6	1/6
7	–	–	–	–	–	–	0	0	0	0
8	–	–	–	–	–	–	–	0	0	0
9	–	–	–	–	–	–	–	–	0	0
10	–	–	–	–	–	–	–	–	–	0

The probabilities do not depend on  $R_0$

5.2.1 The importance of the length of the observed chain

It should be observed that when calculating the probabilities  $P_{ij}^\nu$  we have conditioned on the event that there is a transmission chain of at least length  $\nu$ . If we only wish to calculate the probabilities for the  $\nu_1 < \nu$  first infections and only condition on the event that at least  $\nu_1$  infections have been observed we will get other probabilities. The same is of course true if we know more infections occur after time  $\tau_\nu$ . To illustrate this we will reanalyze the same situation as in Sect. 5.2 using only the four first infections. The estimated probabilities are given in Table 5

**Table 4** Estimated expected number of secondary infections due to  $i$ 'th infected

Case	$R_0$	Infector									
		1	2	3	4	5	6	7	8	9	10
1	1.2	3.13	2.01	1.86	1.02	0.61	0.61	0.44	0.20	0.10	0.01
1	2	2.88	2.03	1.94	1.08	0.64	0.61	0.46	0.22	0.10	0.01
1	4	2.23	2.02	1.40	1.18	0.66	0.66	0.48	0.25	0.11	0.02
2	1.2	3.59	2.39	2.20	0.68	0.48	0.43	0.22	0.01	0.00	0.00
2	2	3.54	2.40	2.23	0.68	0.48	0.43	0.22	0.01	0.00	0.00
2	4	3.38	2.49	2.34	0.68	0.46	0.43	0.20	0.01	0.00	0.00
3	–	4.25	2.25	2.25	0.58	0.33	0.33	0	0	0	0

**Table 5** Estimated probabilities that the  $i$ 'th infected infected the  $j$ 'th infected, using the model as in case 1 with  $R_0 = 2$

i	j		
	2	3	4
1	1	0.61	0.24
2	–	0.39	0.34
3	–	–	0.41

**Table 6** Estimated probabilities for possible chains with four infected using the model as in case 1 with  $R_0 = 2$

Chain	Probabilities
(–,1,1,1)	0.20
(–,1,1,2)	0.18
(–,1,1,3)	0.23
( $i_1, i_2, i_3, i_4$ ) list the infectors in time order	(–,1,2,1) 0.05
– Denotes an infector from outside	(–,1,2,2) 0.16
	(–,1,2,3) 0.18

If there is no individual variation, as in case 3, the probabilities can be read from the upper left-hand corner of Table 3.

5.2.2 Simultaneous probabilities

We have calculated marginal probabilities that the  $i$ 'th infected infected the  $j$ 'th. The events that the  $i$ 'th infected infected both the  $j$ 'th and the  $i^{**}$ 'th infected infected the  $j^{**}$ 'th are not independent. Thus the simultaneous probability is in general not equal to the product of the corresponding marginal probabilities. However, this will be true if there is no individual variation, as in case 3. To illustrate this we have calculated the probabilities for the six possible chains in the example treated in the previous section with four infected in the tree. These probabilities are given in Table 6.

For trees with more observations the number of possibilities will soon get overwhelmingly large. For the chain with  $\nu = 11$  there are  $10!$  possible trees to consider, most of them with very small probabilities.

Observe that the expected number of secondary infections depends on the marginal probabilities only.

## 6 Discussion

Any stochastic analysis of data on transmission of infections will have to depend on some more or less complex model. It is necessary to judge which sources of variation can influence the conclusions and therefore should be taken into account. The purpose of this paper is to illustrate how individual variability in latent and infectious times can influence the probability that an infection is transmitted from a particular infective person. By analyzing a special and very simple model we have illustrated that such individual variations are important. This will also be the case in more complicated and realistic models. The effect will be larger the more individual variation there is. In a particular study individual variability in infectivity may or may not be important depending on the purpose of the study. The possibility that important conclusions can be influenced are worth to be considered.

We have considered two very different situations. In the first we investigate where an infection was transmitted under the assumption that there is no connection between the potential infectors. In the second situation we consider that we have a tree of infections and try to calculate the (marginal) probabilities for different transmission links. In both situations we have conditioned on the assumed observation scheme. The conditioning has consequences. It is shown in the numerical examples that the length of the observed transmission tree influences the probabilities. This is reasonable, since if we know that many individuals in a population eventually are infected we can conclude that the early infectors will probably have been powerful transmitters. If they were not, it would be likely that the infection tree would have stopped early.

We have throughout assumed homogeneous mixing, in the sense that there is no prior differentiation between transmission links, i.e., all possible pairs of persons can be involved in a transmission. The only information that used is the times of infections. In real cases other information may be available, such as family connections, spatial or social closeness. Another interesting possibility would be to use data on genetic diversity of the infectious agent (cf Jombart et al. 2011). There may also be other information such as all infections in an outbreak are observed. If that is the case also other parameters of interest may be estimated, e.g.  $R_0$ .

A more surprising result may be that in a transmission tree not only the timing of infections but also the strength of infectivity measured by e.g.  $R_0$  is important. This is the case if there is individual variation in infectivity. Heuristically we can understand this as an effect of the conditioning. If the infections are sparse and  $R_0$  is high it can be explained by short infectious times.

Exact times are often difficult to observe. It may be of interest to consider other times related to the infections. Removal times are the times when the infected stops



being infectious. An analysis of transmission routes based on observations of removal times will be complicated by the possibility that an infected person may be removed before the removal of its infector. Another possibility is that infections are recorded when the first symptoms occur. For simplicity we can assume that the occurrence of symptoms coincides with the start of the infectious period. This implies that an infected is observed at the end of the latent period. In this case the distance between the observation times of the infector and the infected are related to the latent time of the infected person. The distance between infection times are related to the latent time of the infector. The analysis above is built on the random event that the  $i$ 'th infector is infectious at the time the  $j$ 'th infected is infected. Let  $\sigma_1, \dots, \sigma_v$  be the observed times of first symptoms and let the indicator  $\tilde{I}_i(\sigma_j - \sigma_i)$  equal 1 if the  $i$ 'th observed is infectious at the time the  $j$ 'th observed is infected and 0 if not. If  $L_j$  is the latent time of the  $j$ 'th observed and  $a = \sigma_j - \sigma_i$  then

$$Pr(\tilde{I}_i(a) = 1 \mid L_j) = 1 - H_I(a - L_j). \quad (17)$$

Thus  $Pr(\tilde{I}_i(a) = 1) = \int_0^a (1 - H_I(a - s))h_L(s)ds$ . If the duration of the latent and infectious times are independent (for the same individual), as is assumed in the three cases considered above, this probability equals the probability  $p(a)$  given by Eq. (9). However, when calculating the expectation of  $Z_i$  (cf Eq. 12) or  $Z_{ij}$  (cf Eq. 15) we can no longer assume that the indicators involved are independent. This will require other methods of calculations resulting in other probabilities. However, we expect that the qualitative results of an analysis based on times of occurrence of symptoms will be similar to the results obtained above.

## References

- Bremaud P (1981) Point processes and queues. Martingale dynamics. Springer, New York, Heidelberg, Berlin
- Carrat F, Vergu E, Ferguson NM, Lemaître M, Cauchemez S, Leach S, Valleron AJ (2008) Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *Am J Epidemiol* 167(7):775–785
- Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM, Swerdlow W, The Pennsylvania H1N1 working group (2011) Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *PNAS* 108(7):2825–2830
- Cauchemez S, Ferguson NM (2012) Methods to infer transmission risk factors in complex outbreak data. *J R Soc Interf* 9(68):456–469
- Hens N, Calatayud L, Kurkela S, Tamme T, Wallinga J (2012) Robust reconstruction and analysis of outbreak data: Influenza A(H1N1)v. Transmission in a school-based population. *Am J Epidemiol* 176(3):196–203
- Jombart T, Eggo RM, Dodd PJ, Balloux F (2011) Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* 106:383–390
- Kenah E (2011) Contact intervals, survival analysis of epidemic data, and estimation of  $R_0$ . *Biostatistics* 12(3):548–566
- Svensson Å (2007) A note on generation times in epidemic models. *Math Biosci* 208(1):300–311
- Tomba GS, Svensson Å, Asikainen T, Giesecke J (2010) Some model based considerations on observing generation times for communicable diseases. *Math Biosci* 223(1):24–31
- Wallinga J, Teunis P (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* 160(6):509–516