Infectious disease transmission as a forensic problem: who infected whom?

Peter Teunis1,2, Janneke C. M. Heijne3, Faizel Sukhrie4,5, Jan van Eijkeren6, Marion Koopmans4,5 and Mirjam Kretzschmar1,7

1Epidemiology and Surveillance Unit, 2Laboratory for Infectious Diseases and Perinatal Screening, Centre for Infectious Disease Control, and 4Expertise Centre for Methodology and Information Services, RIVM, Bilthoven, The Netherlands
2Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA
3Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
4Department of Virology, Erasmus Medical Centre, Rotterdam, The Netherlands
5Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, The Netherlands

Observations on infectious diseases often consist of a sample of cases, distinguished by symptoms, and other characteristics, such as onset dates, spatial locations, genetic sequence of the pathogen and/or physiological and clinical data. Cases are often clustered, in space and time, suggesting that they are connected. By defining kernel functions for pairwise analysis of cases, a matrix of transmission probabilities can be estimated. We set up a Bayesian framework to integrate various sources of information to estimate the transmission network. The method is illustrated by analysing data from a multi-year study (2002–2007) of nosocomial outbreaks of norovirus in a large university hospital in the Netherlands. The study included 264 cases, the norovirus genotype was known in approximately 60 per cent of the patients. Combining all the available data allowed likely identification of individual transmission links between most of the cases (72%). This illustrates that the proposed method can be used to accurately reconstruct transmission networks, enhancing our understanding of outbreak dynamics and possibly leading to new insights into how to prevent outbreaks.

1. Introduction

Studies of infectious disease transmission are often based on simulations, for instance of the possible outcomes when an infectious subject is released into a susceptible population. Careful consideration of the probabilities of any possible outcome then may be used for statistical inference on important parameters, for transmissibility and mixing behaviour [1–4]. When there is knowledge about contacts between individuals, simulations may incorporate potential paths of the infection in a given network [5–8]. A complication is that transmission of infection depends not only on who is connected to whom, but also on when (in which order) these contacts occurred [9–11]. For any but the smallest populations with simple structure, a forward projecting approach is problematic, because of the many different outcomes that are possible.

However, in a given population of cases as may be observed in an outbreak of infectious disease, reversing time makes the problem a lot simpler: any infected subject can have been caused by only one infectious ancestor [12–15]. Exploiting this notion, our goal is to track possible links between cases of infectious disease, identifying who may have been infected by whom. In order to do so, we want to use all relevant information between each pair of cases. Information about a possible causal link between a pair of cases may be found in the times of their occurrence, their spatial locations, in their serological status or possible differences in genetic sequences of the infecting pathogen. Information from each of these sources can be interpreted as a measure of the disparity between pairs of cases, similar properties indicating proximity. The greater the dissimilarity, the
greater the distance between cases, in terms of time, location, serum antibody titre or genetic sequence of the pathogen.

When an outbreak is studied, various categories of data are collected, but usually some information is missing. Records from any individual patient may miss one or more items, and few records may be complete. The approach presented here combines various incomplete data sources to estimate who infected whom as accurately as possible.

To outline the approach used here, we will start by defining a transmission matrix of the probabilities that transmission occurred between any pair of cases in the affected population. Then we will show how to formulate a likelihood function, allowing estimation of the elements of this matrix. The likelihood contains a pairwise kernel function dependent on the available data mentioned above (time, location, …) and we will show how these various data sources may be incorporated. Finally, use of the estimated transmission matrix to estimate any subject must have acquired their infection from exactly one other subject in the observed infected population. In case of an incompletely observed outbreak, e.g. asymptomatic cases contributing to transmission, this latter condition may have to be relaxed.

By characterizing links through their parent nodes, rows can be estimated independently: the probability that subject \( i \) obtained its infection from subject \( j \) is independent of the probability that subject \( k \) obtained its infection from subject \( j \).

Note that the elements \( v_{ij} \) are probabilities of transmission that do not directly identify links between subjects. A transmission link for subject \( i \) may be thought of as a random sample from a categorical distribution with probability vector \( n \).

2.2. Kernel likelihood
Suppose observations of a distance measure \( X_{ij} \) can be made on pairs of subjects \((i,j)\). These distances may be serial intervals for symptom onset, or spatial distance, or any other quantifiable measure that may indicate a causal link between two subjects.

Suppose \( \kappa_{ij}(X_{ij}|i\leftarrow j;\theta) \) is the conditional probability (density) of \( X_{ij} \) given that subject \( i \) was infected by \( j \) (designated as \( i\leftarrow j \)). If \( \kappa_{ij}(\cdot) \) is a parametric function, it may depend on parameters \( \theta \); these will be omitted unless they are needed. The contribution of the pair \((i,j)\) to the likelihood then is

\[
\ell_{ij}(X_{ij}, v_{ij}) = \kappa_{ij}(X_{ij}|i\leftarrow j)v_{ij}. \tag{2.4}
\]

Note that if subject \( i \) is known not to have had contact with subject \( j \) the probability of transmission between them is \( v_{ij} = 0 \). Omitting constant (normalizing) terms, the dot product

\[
\ell_i(X_i, v_i) = \kappa_i(X_i) \cdot v_i = \sum_{j=1}^{n} \kappa_{ij}(X_{ij}|i\leftarrow j)v_{ij} \tag{2.5}
\]

can be interpreted as a likelihood function, with probabilities of transmission \( v_i \) from any possible ancestors to case \( i \), given the observed data \( X \).

Over all observed subjects, the likelihood becomes

\[
\ell(X, V) = \prod_{i=1}^{n} \ell(X_i, v_i) \propto \prod_{i=1}^{n} \kappa_i(X_i) \cdot v_i. \tag{2.6}
\]

2.3. Estimating \( V \)
Given a set of observed distances \( X \) between pairs of cases, the transmission probability matrix \( V \) can be estimated, using the likelihood function in equation (2.6).

As any row \( v_i \) of \( V \) represents the vector of probabilities that case \( i \) received their infection from any other case, these row vectors may be estimated independently of one another.

Finding the maximum likelihood link to case \( i \) is trivial: this is a vector with \( v_{ik} = 1 \) at the position \( k \) of the largest element of \( v_i \) and 0 elsewhere (see the electronic supplementary material, appendix D). If all elements in \( v_i \) are close to 0 and only \( v_{ik} \approx 1 \) this is sensible; however, if all elements in \( v_i \) were to be approximately equal with only one (at position \( k \)) slightly larger than the rest, the maximum likelihood vector would also have \( v_{ik} = 1 \) and 0 elsewhere (see the electronic supplementary material, appendix D), but we would be less certain of who was the ancestor of case \( i \), than in the former case.
To obtain information about the uncertainty in the transmission probabilities $v_{ij}$ they can be estimated in a Markov chain Monte Carlo (MCMC) procedure. Using a Metropolis–Hastings algorithm [16] we start with an initial matrix $V$ with arbitrary elements (for instance all elements in any row not known to be zero set to the same value). The likelihood function (equation (2.6)) is then used to calculate the likelihood of that initial position in the Markov chain. In case prior distributions are specified, a posterior probability (density) may be calculated.

In order to allow use of a normal candidate distribution, the $v_{ij}$ are first logit transformed,

$$z_{ij} = \log \left( \frac{v_{ij}}{1 - v_{ij}} \right)$$

(2.7)

and candidates $z'_{ij} = z_{ij} + \epsilon$ are calculated ($\epsilon \sim \text{N}(0,\sigma)$), and the new $v'_{ij}$ are obtained by reverse logit transform and scaling so that rows add to 1

$$u'_{ij} = \frac{e^{v'_{ij}}}{1 + e^{v'_{ij}}}$$

and

$$v'_{ij} = \frac{u'_{ij}}{\sum_{j=1}^{k} u'_{ij}}.$$  

(2.8)

The likelihood for the candidate matrix is then calculated (or its posterior), and the candidate is accepted, if ratio of the likelihoods (or posteriors) for $V$ and $V$ is greater than a uniform (0,1) random variate [16].

Note that here the kernel function $\kappa_{ij}(X_i,|i \rightarrow j,\theta)$ has been assumed known, that is: $V$ is estimated, given the parameter set $\theta$. Ultimately, we need to infer both $V$ and $\theta$ (see the electronic supplementary material, appendix E).

### 3. Linking kernel function

Information to identify pairs of cases as links in a transmission chain may be related to the infection, such as the time between symptom onsets or the (genetic sequence of the) pathogen involved. Other information may relate to contact patterns that exist independent of the outbreak, such as spatial location of the subjects or contact frequencies between subjects (though these latter may be modified by disease).

Observations may take the form of a set of numbers representing distance in some trait space, where distance is associated with a probability of transmission. The matrix of $\kappa_{ij}(X_i,|i \rightarrow j,\theta)$ then may have elements consisting of parametric probability densities with some common set of parameters $\theta$.

Often, such distance measurements are not available, but there may be knowledge of the probabilities of links between subjects that can be quantified. For instance, two subjects may or may not share the same peer group, implying frequent or less frequent contacts, or two subjects may be infected with genetically distinct isolates of the same pathogen sufficiently different to exclude mutation. Such data can be incorporated by building a matrix of fixed elements $\kappa_{ij}$. In the following, we specify different types of distance measures.

#### 3.1. Serial interval

In the previous studies, the likelihood of a transmission link between two cases has been calculated by way of the serial interval distribution [12]. The serial interval is defined as the time between successive clinical cases [17].

When the hazard function for onset of illness in a secondary case, given the interval $\tau$ between onsets of illness symptoms is $h(\tau|\theta)$ with model parameter vector $\theta$, and we would know the onsets $T_i$ any two cases $i,j$ separated by a serial interval $T_i - T_j$ would be linked [18] by a probability

$$\kappa_{ij}^T(T_i, T_j|\theta) = \frac{h(T_i - T_j|\theta)}{\sum_{k=1}^{n} h(T_i - T_k|\theta)}.$$  

(3.1)

When recorded times are interval censored, for instance in days, the kernel function must be slightly modified (electronic supplementary material, appendix A).

In practice there are two additional constraints that may save computational effort;

— negative serial intervals are excluded. When infected subjects become infectious long before they start having symptoms, the incubation period in one subject can be shorter than the interval between onset of infectiousness and onset of symptoms in its ancestor [17], resulting in a negative serial interval (figure 1). For most diseases this is not likely and for convenience it is assumed that any element with $T_j > T_i$ must have a kernel likelihood zero. If needed, this assumption may be easily dropped; and

— when the serial interval between two cases exceeds the maximum serial interval (say, some high percentile of its distribution), a link is highly unlikely and the corresponding kernel likelihood may also be set to zero.

#### 3.2. Proximity of cases

The distances between cases can be expressed as geometrical distances, calculated from known geographical locations, but

---

**Figure 1.** The serial interval between successive cases in a transmission chain: the generation interval $T_g(i,j)$; the serial interval for onset of infectiousness $\Delta T_{\text{inf}}(i,j)$ and the serial interval for symptom onset $\Delta T_{\text{sym}}(i,j)$. When the incubation period (from infection to symptom onset) is long, negative serial intervals for symptom onset may be found.
alternatively, some more or less precise proximity measure may be given. When the locations of cases (e.g. home or work) are given such continuous distance measures may be evaluated in a contact kernel, similar to the generation interval kernel above [19,20].

However, proximity of any pair of cases may often be more appropriately characterized by a measure of social distance, reflecting the probability of contact between two cases. For instance, membership of a peer group increases the probability of contact and not being part of that group decreases the probability of contact. The probabilities of any pair of cases sharing contacts then define a (social) proximity kernel

\[ k_{ij}^{s} = \begin{cases} a_0 & i \text{ and } j \text{ in the same group} \\ a_1 & i \text{ and } j \text{ in different groups} \end{cases} \tag{3.2} \]

where the two parameters \( a_0 \) and \( a_1 \) may be probabilities or rates of contacts, the ratio of their magnitudes representing the ratio of contact probabilities between subjects.

3.3. Genetic similarity

When pathogen samples have been taken from patients, genetic sequence information may be available. Such sequences are commonly collected for constructing the genetic lineage of the pathogen. However, the genetic similarity between any pair of isolates from the same cluster of infectious disease cases may also be used as a measure of the distance between the cases the samples were obtained from. Genetic similarity between pathogens from a pair of cases \( i \) and \( j \) can be expressed as a genetic kernel coefficient \( G_{ij} \) weighting links between observed cases [21]. Some methods for characterizing genetic distance discriminate between \( G_{ij} \) and \( G_{ij',o} \) others not.

Often simpler, categorical information is available. Suppose a limited number of different genotypes of a pathogen may be found, that are sufficiently different so that one type cannot transform into another in a few generations, within an outbreak. Then a cluster of cases with mixed genotypes must consist of a mixture of transmission networks, one for each genotype. Any case may shed pathogens of any one of \( k = (1, \ldots, K) \) different genotypes (assuming mixed infections do not occur). Pathogen type can be represented by a vector \( c_i \) of length \( K \). When case \( i \) sheds pathogen type \( k \) then all elements of \( c_i \) are zero, except the element at position \( k \) which is 1. The genetic kernel function

\[ k_{ij}^{G} = c_i \cdot c_j \tag{3.3} \]

now leads to a matrix with elements 0 or 1, for cases with either the same or different genotypes.

When the genotype of the pathogen has not been observed in all cases, the abundance of observed genotypes may be used to assign a probability that a case has any one of the set of possible types. The distribution of the abundance of each genotype may be assumed multi-nomial with probability vector

\[ p = \{p_1, p_2, \ldots, p_k\} \tag{3.4} \]

with element \( p_k \) as the probability that any unknown specimen from the outbreak has genotype \( k \) and the sum of all \( p_k \) adding up to 1. Assuming equal detectability of genotypes, these abundances (the probability that any detected virus is of a particular genotype) may be obtained from the observed frequencies in the same outbreak.

Genotypes may be assigned to the cases with missing genotype by drawing a random sample of the abundance distribution. For any single case, this is a sample from a categorical distribution (multi-nomial with \( n = 1 \)). For any specific case \( i \), the different possible genotypes produce different optimum (maximum likelihood) estimates of its links to other cases. Assigning a different genotype to a case where that information is missing is equivalent to assigning that case to a different transmission sub-network.

3.4. Combining different linking kernel functions

Analogously, other types of information on distances between cases may be formulated in terms of kernel functions. The information from onset times, locations and/or genetic similarity may be combined to produce a likelihood contribution for each possible pair of observed cases. In the simplest case of independent contributions, the elements \( k_{ij} \) of the likelihood matrix then are found by element-wise multiplication

\[ k_{ij} = k_{ij}^{s} k_{ij}^{G} k_{ij}^{X} \tag{3.5} \]

assuming absence of correlation between types of distances (time, location and genetic sequence). The superindexes denote kernels based on time \((T)\), spatial or social \((S)\) and genetic \((G)\) data, respectively.

4. Analysis of the estimated transmission probability matrix

The transmission probability matrix provides information about the transmission network of an outbreak on several levels: distribution of reproduction numbers and other network metrics, distribution of possible transmission trees, and identifiability of specific links between cases. The set of \( V \) matrices obtained by MCMC sampling may be analysed for studying these properties of the transmission network.

4.1. Reproduction number

Row \( i \) of the transmission matrix shows a vector of the probabilities of case \( i \) to have acquired their infection from any other case. Conversely, column \( j \) of this same matrix is a vector of the probabilities of case \( j \) to have transmitted their infection to any other case. Their sum

\[ R_j = \sum_{i=1}^{n} v_{ij} \tag{4.1} \]

can be interpreted as the expectation of the number of cases produced by infectious case \( j \), or its reproduction number. Thus, we can use the transmission matrix to calculate reproduction numbers for any individual case in the outbreak. In network terminology, \( R_j \) is an estimate of the outdegree of node \( j \) [22].

The posterior sample of transmission probability matrices (obtained from the Markov chain) can be considered a sample from the joint posterior distribution of transmission probabilities for all cases. If an element from the iteration with rank number \( m = 1, \ldots, M \) is \( v_{ij}^{(m)} \), then the sample is

\[ R_j = \left\{ \sum_{m=1}^{M} v_{ij}^{(1)} \sum_{m=1}^{M} v_{ij}^{(2)} \ldots \sum_{m=1}^{M} v_{ij}^{(M)} \right\} \tag{4.2} \]
Quantiles from such a sample of $R_i$ may be used to calculate marginal posterior intervals of the reproduction numbers. Using the transmission probability matrix, other network metrics can also be calculated.

### 4.2. Transmission tree

As mentioned above (§2.1), links between case $i$ and any other case can be assumed realizations of a random sample from a categorial distribution with probability vector $v_{ij}$. Such sampling reduces each row of $V$ to a binary vector with exactly one element 1 and the rest zero. This reduced matrix may be interpreted as an adjacency matrix, equivalent to a directed graph representing a transmission tree of the outbreak.

Note that in order for $V$ to be translated into a proper transmission network, cycles are not admissible. If any element $v_{ij} \neq 0$, then $v_{ji} = 0$: subjects cannot mutually infect each other. Any larger cycles (i.e. loops involving more than two subjects) also cannot occur. By sorting all cases in order of their onsets of symptoms and preventing negative serial intervals (and treating degenerate cases as in the electronic supplementary material, appendix B), cycles are prevented from occurring.

### 4.3. Identifiability of links

Any row $i$ of the transmission probability matrix $V$ represents a vector of probabilities that a subject $i$ got their infection from any of the other cases in the outbreak. If that vector has a single non-zero element (that then must be 1, if this is a closed outbreak), then the person who infected subject $i$ can be identified uniquely. In reality, this will rarely happen: several elements of the vector at row $i$ can be non-zero, pointing to several other subjects as possible infectors of subject $i$.

To study the ambiguity regarding who infected subject $i$, a crude measure might be the number of non-zero elements in row $i$ of the transmission matrix. However, it is not likely that all of these possible candidates were assigned the same probability of having infected subject $i$.

The probability that $i$ was infected by another infectious subject $j$ and not by anyone else is $v_{ij}$. Interpreting infection of $i$ by $j$ as a Bernoulli trial with variance

$$\text{var}_{ij} = v_{ij}(1 - v_{ij})$$

may be used as a measure of how accurately $j$ may be identified as the person who infected $i$. The sum

$$q(i) = \sum_{j \neq i} \text{var}_{ij} = \sum_{j} v_{ij}(1 - v_{ij})$$

then is a measure of how accurately any of the subjects in the observed population may be identified as having infected subject $i$.

The quantity $1 - q(i)$ may be used as a score of the degree of identification of the infection parent of subject $i$. When there is only one non-zero element in row $i$ of the transmission matrix, the parent node is perfectly known, and $1 - q(i) = 1$ indicating perfect identification. When row $i$ has $n_i$ non-zero terms and they all have the same value $1/n_i$ then

$$1 - q(i) = 1 - \sum_{i=1}^{n_i} \frac{1}{n_i} \left(1 - \frac{1}{n_i}\right) = \frac{1}{n_i}$$

is at a minimum. This would represent complete ignorance about who infected subject $i$, among those for whom a link should be possible (i.e. those with non-zero $v_{ij}$).

### 5. An application: nosocomial transmission of norovirus

Norovirus is an enteric pathogen causing a considerable proportion of outbreaks of enteric illness worldwide [23], in particular in hospitals and nursing homes [24]. Norovirus can be easily transmitted by infectious cases [13,14]. Various different genotypes of the infectious pathogen may be isolated from patients in norovirus outbreaks. Such diversity may result from exposure to a common source containing a mixture of virus genotypes [25]. However, in outbreaks with person-to-person transmission [26], such a cluster must be considered a mixture of as many outbreaks as there are identified genotypes (§3.3).

This analysis uses previously published data from norovirus outbreaks that occurred during a 5 year period (19 December 2002–31 December 2007) in a university hospital in the Netherlands [27]. In short, from 264 cases in total, 160 provided samples that were analysed (39.4% missing). Data included only cases with symptoms of acute gastroenteritis (vomiting, diarrhea), as date of symptom onset. The hospital department where any of these cases occurred was also known. Assuming that cases from different departments had a decreased probability of transmission owing to the spatial separation, we set a fixed weight factor for transmission between cases in the same department (0.9) and between cases from different departments (0.1), similar to previously reported analyses [13,14].

It is a rare occasion that all cases in an outbreak are sampled for shedded pathogens. For that reason, when multiple genotypes have been identified in part of the identified cases but a fraction (with similar symptoms) lacks samples and remains untyped, these untyped cases may be assigned to any of the networks describing transmission for each genotype (§3.3).

As the elements of the transmission probability matrix are updated in the MCMC procedure, randomly rotating all available genotypes over all cases with missing genotype results in large steps in posterior probability space. Therefore, a nested procedure was used where genotypes of random untyped cases are updated one at a time, followed by updating the transmission probability matrix for several thousands of iterations. By this procedure, each case with missing genotype is inserted into the transmission network, where it causes the strongest increase in posterior probability.

Norovirus may be transmitted through direct contact between an infectious and a susceptible person, but often the virus remains infectious for extended periods when deposited on inanimate surfaces. Anyone touching such contaminated surfaces may then transfer some infectious virus to the skin of their hands (for instance) and subsequently inoculate themselves by touching exposed mucosal areas. Such an environmental transmission route may link cases over extended intervals. The serial interval may, therefore, be heterogeneous: in addition to the short-term (relatively) high probability of transmission by direct contact, environmental transmission causes a tail effect with a low frequency of long serial intervals (figure 2).
Figure 2. Serial interval distribution as a binary mixture showing two modes of transmission: a short interval corresponding to direct contact [13,28], and a more disperse, longer interval corresponding to environmental transmission.

Figure 3. Frequencies of the different norovirus genotypes found in a multi-year hospital outbreak study (obs) compared with frequencies assigned by transmission analysis, sampled from a Markov chain of their posterior distribution: median and 95% intervals.

Figure 4. Example transmission trees based on transmission matrix estimates for the two most common genotypes, GII.3 and GII.4. Letters indicate (hospital) departments, and cases with missing genotype are indicated with an asterisk (*).

Table 1. Parameters and quantiles of the binary mixture used for the temporal kernel distribution (figure 2). Shown are the weight factor for each component (proportion), the shape and scale parameters of the gamma distributed components, and their arithmetic means and a 95% range.

<table>
<thead>
<tr>
<th>component</th>
<th>proportion</th>
<th>shape</th>
<th>scale</th>
<th>mean</th>
<th>Q₀.₀₂₅</th>
<th>Q₀.₉₇₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.306</td>
<td>6.53</td>
<td>0.31</td>
<td>2.00</td>
<td>0.77</td>
<td>3.80</td>
</tr>
<tr>
<td>2</td>
<td>0.694</td>
<td>0.88</td>
<td>18.07</td>
<td>8.73</td>
<td>2.23</td>
<td>19.56</td>
</tr>
</tbody>
</table>
In this analysis, a distribution for the serial interval was determined as a mixture of the distribution obtained from day-care centre data [13, 28] and a component with a similar shape (gamma distribution) but longer intervals, representing possible environmental transmission. Note that there is still only one transmission probability matrix, but its elements now represent a mixture of two modes of transmission. Initial parameters for this distribution (table 1) were determined using the unweighted kernel function described in the electronic supplementary material, appendix C. As soon as convergence was established, the parameters of this serial interval kernel were also estimated in the MCMC procedure (see the electronic supplementary material, appendix E). Figure 2 shows the shape of the temporal kernel function used in this analysis.

When the cases with ‘missing’ genotype were included into the transmission analysis, using their symptom onset and hospital department data, a likely genotype could be inferred. Figure 3 shows the distribution of norovirus genotypes in observed patient samples, and the estimated (posterior) frequencies of these genotypes.

Figure 4 shows examples of transmission trees, based on the transmission matrix with the highest posterior probability (‘plausible estimate’), for the most frequent genotypes GII.3 and GII.4.

Now we ask whether the analysis led to conclusive results on who was infected by whom? Using the score metric defined above parent nodes could be identified with varying degrees of certainty. Some nodes remained isolated \( \left( \frac{N}{\text{node}} = 16\% \right) \) meaning that no parent nodes were found, because there were no eligible cases within the time window defined by the temporal kernel distribution (figure 2). These may have been founders infected by a source outside the observed cluster, but they could also have been infected by an undetected (asymptomatic) case. For the remaining cases, it was frequently possible to identify a parent node from whom they received their infection, summarized in table 2. For about 18 per cent of all cases \( \left( \frac{N}{\text{node}} \right) \), a unique parent node could be identified \( \left( 1 - q = 1 \right) \), and for another 12 per cent of cases, a parent node could be found with only minor uncertainty \( \left( 31\% \right) \). For 24 per cent of cases parent nodes could not be identified \( \left( 21\% \right) \). Figure 5a shows the distribution of scores for all nodes. Compared with nodes representing subjects with known virus genotype, nodes with unknown genotype show greater variability in their identifiability scores (figure 5b). Figure 5c shows how nodes with a greater possible number of parents tend to have lower scores, as may be expected. Typical results for the four categories of outcomes (complete certainty, minor uncertainty, substantial uncertainty and complete ignorance) are shown in the electronic supplementary material, appendix A4.

Of course the identification of parent nodes still rests on the assumption that infectious cases in these outbreaks were all observed. While it is likely that cases identified as isolated have been infected outside the hospital, it remains uncertain whether any cases in the observed outbreaks may have been infected by unobserved contacts from outside the hospital, or indeed by asymptomatic subjects who remained invisible to the outbreak detection protocol [15]. The missing data approach, used here to deal with genotype data, can be
extended to study whether addition of a missed infectious subject may substantially improve the posterior probability of the whole transmission network.

6. Conclusion

The present paper has shown that by combining observed data from different modalities in a pairwise analysis of all cases in an outbreak, the question of ‘who has infected whom’ may be answered, for a majority of the patients involved in an outbreak. Such detailed information on the transmission of an infectious disease is valuable, to identify factors that enhance or impair transmission, or to identify superspreaders.

The transmission probability matrix that is obtained also defines a transmission network (including uncertainty in the links connecting nodes), providing information about the contact network, the underlying infrastructure on which the disease is transmitted. The methods used here may support the analyses of transmission networks, based on observed infectious disease data.

We thank much to our colleagues Simon Heisterkamp, Nico Någellkerke, Hans van Houwelingen and Jacco Wallinga, whose critical advice in many discussions has greatly helped to improve the paper.

References