Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data

R. J. F. Ypma^{1,2,*}, A. M. A. Bataille^{3,4}, A. Stegeman³, G. Koch⁴,
J. Wallinga^{1,2} and W. M. van Ballegooijen¹

¹National Institute of Public Health and the Environment, Bilthoven, The Netherlands

²Julius Centre for Health Research and Primary Care, University Medical Centre, Utrecht, The Netherlands

³Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University,
Utrecht, The Netherlands

⁴Department of Virology, Central Veterinary Institute, Animal Sciences Group, Wageningen University
and Research Centre, Lelystad, The Netherlands

Knowledge on the transmission tree of an epidemic can provide valuable insights into disease dynamics. The transmission tree can be reconstructed by analysing either detailed epidemiological data (e.g. contact tracing) or, if sufficient genetic diversity accumulates over the course of the epidemic, genetic data of the pathogen. We present a likelihood-based framework to integrate these two data types, estimating probabilities of infection by taking weighted averages over the set of possible transmission trees. We test the approach by applying it to temporal, geographical and genetic data on the 241 poultry farms infected in an epidemic of avian influenza A (H7N7) in The Netherlands in 2003. We show that the combined approach estimates the transmission tree with higher correctness and resolution than analyses based on genetic or epidemiological data alone. Furthermore, the estimated tree reveals the relative infectiousness of farms of different types and sizes.

Keywords: transmission tree; molecular epidemiology; avian influenza

1. INTRODUCTION

Estimating the transmission tree for an epidemic of an infectious disease can provide valuable insights; it has been used to evaluate the effectiveness of intervention measures [1–4], to quantify superspreading [5], to identify mechanisms of transmission [6] and to study viral evolutionary patterns [7,8].

Unfortunately, estimating the transmission tree is rarely trivial. Generally, one needs detailed epidemiological data, such as contact structures, while for many epidemics only general statistics, such as case lists with time of symptom onset, are known. Furthermore, data are often missing for some cases. Luckily, another valuable source of information, genetic data on the pathogen, is becoming increasingly available. The amount of genetic diversity observed between samples taken from different cases informs us of the distance in the transmission tree between these cases.

Estimates of the transmission tree will be best when all available data are combined in one analysis; however, methods to achieve this are still largely lacking [9]. One approach proposed by Cottam *et al.* [10] is to use genetic data to exclude certain potential transmission trees, and then evaluate the remaining possible trees

with epidemiological data. Jombart *et al.* [11] followed a similar approach, contrasting the construction of transmission trees with the construction of phylogenetic trees.

We argue that a more consistent approach can be obtained by combining both genetic and epidemiological data in one likelihood function when reconstructing the transmission tree. This likelihood function can then be used in a Bayesian setting to sample from the space of all transmission trees, allowing for the simultaneous estimation of both the tree and the parameters of the function itself. These parameters themselves can be used to describe the dynamics of the epidemic. Such an approach would be able to handle missing data, e.g. cases for which no genetic data are known.

We develop the approach to obtain a more detailed understanding of an epidemic of avian influenza A (H7N7) in different types of poultry farms in The Netherlands in 2003. This epidemic spread over a large area and infected 241 farms and 89 humans in total, with one human fatality [12–14], even though a movement ban was promptly imposed on confirmation and both infected and suspected farms were culled. The epidemic also spread abroad, infecting eight farms in Belgium and one in Germany. This specific H7N7 strain has not been detected since, indicating the epidemic formed a dead end for virus spread.

Using only epidemiological data, previous studies showed there is a strong spatial component to spread [15] and small hobby farms are less susceptible to infection than large commercial farms [16]. However, it

* Author for correspondence (rolf.ypma@rivm.nl).

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rspb.2011.0913> or via <http://rspb.royalsocietypublishing.org>.

remains unclear how the virus spread from farm to farm or how farm size and type relate to infectiousness. Previous studies for other farm animal diseases such as foot-and-mouth disease found a clear relation between the number and type of animals on a farm and their infectiousness [1,2,17]. We might expect similar differences in infectiousness for avian influenza H7N7 as well, for example, because the transport structures are different for different farm types (e.g. food transports and egg rearing) and the different host species react differently to the virus. Here, we quantify the infectiousness of farms of different types and sizes, using transmission trees that we reconstructed from genetic, geographical and temporal data.

2. METHODS

(a) Data

Influenza A (H7N7) virus was detected on a total of 241 farms in The Netherlands. These consisted of 205 commercial chicken farms, 14 hobby farms (defined as flocks of less than 300 animals), 19 turkey farms and three duck farms. The nine farms infected abroad have not been included in this analysis because virus sequences of these outbreaks, exact location and/or date of infection were not known to us and the farms are considered dead ends for this epidemic.

For all 241 farms, the geographical location and the date of culling have been recorded. The date of infection has been estimated using animal mortality data for all farms except for hobby farms [18], for which we take the estimates from Boender *et al.* [15]. Results in this paper were found to be robust to small variations in infection dates. Data on the number of animals kept were available for 220 farms. All available data are provided in the electronic supplementary material (although we are only allowed to specify the geographical location in terms of the closest town, owing to privacy restrictions).

For 185 farms, the RNA consensus sequences of the haemagglutinin, neuraminidase and polymerase PB2 genes were determined from a pooled sample of five infected animals [19] (GISAID accession numbers EPI_ISL_68268-68352 and EPI_ISL_82373-82472). We will refer to these farms as sequenced farms, and to the remaining 56 farms as unsequenced farms.

(b) Model

The first farm to be infected in the epidemic is taken as the index case, infected by some unknown source; all other farms are assumed to be infected by a previously infected farm through an unknown route. We make the simplifying assumption that all three types of data are independent from each other, e.g. knowing the geographical distance between a farm and the farm that infected it tells you nothing about the genetic distance between the sequences sampled at those farms or the time between infections of the two farms. Figure 1a illustrates how the approach works. For every farm B, we evaluate which of the previously infected farms could have been the source. The likelihood L that a certain farm A infected B increases if A is not yet culled when B is infected, if A is located close to B, if the sequence taken from farm A is similar to that taken at farm B, and if no other viable candidates exist that could have infected farm B. When genetic information is unavailable for one of

Figure 1. Examples of transmission trees. Circles denote farms, full arrows denote estimated infections, dotted arrows denote possible infections. Filled circles denote farms from which no sequence data were available. (a) When all data are available, the probability that B infected C is proportional to the likelihood $L(\delta_{BC})$ given in equation (2.1). (b) When genetic data are missing for B, we can infer this by looking at the sequence of the farm A that infected B. We then assess the likelihood for the whole subtree: A infecting B, infecting C. (c) When genetic data are missing for multiple farms, we look at the subtrees containing those farms. Here, two transmission trees are possible; B is infected by either A or C. The likelihood that A infected B is proportional to the likelihood of S1, while the likelihood that C infected B is proportional to the product of the likelihoods of S2 and S3.

the farms, we look at the sequence of the farm that infected this farm (figure 1b).

(c) Likelihood without missing data

When there is no missing data, the likelihood of a transmission tree is simply the product of the likelihoods of the links that it consists of. Therefore, we construct a likelihood function L that gives the likelihood for our set of parameters $\mathbf{w} = (b, r_0, \alpha, \mathbf{p})$ and the event δ_{AB} that a certain farm A infected another farm B, given our data D consisting of temporal, geographical and genetic components, $\mathbf{t}$, $\mathbf{x}$ and RNA, respectively. This function consists of a product of contributions given by the temporal, geographical and genetic data:

$$L(\delta_{AB}, \mathbf{w} | D) = L_t(\delta_{AB}, b | \mathbf{t}_A, \mathbf{t}_B) L_{\text{geo}}(\delta_{AB}, r_0, \alpha | \mathbf{x}_A, \mathbf{x}_B) \times L_{\text{gen}}(\delta_{AB}, \mathbf{p} | \text{RNA}_A, \text{RNA}_B). \quad (2.1)$$

(d) Time

We assume that farms are infectious starting 1 day after infection owing to a latent period [18,20,21], and remain equally infectious until they are culled (a sensitivity analysis for the length of the latent period can be found in the electronic supplementary material). Date of infection and culling are denoted by t^{inf} and t^{cull} , respectively. Owing to the decay of contaminated particles, and the possibility of an infection mechanism that introduces time lag, infectiousness drops

exponentially after culling with rate of decline b . This gives us the likelihood of the parameter b and farm A infecting another farm B given their temporal data:

$$L_t(\delta_{AB}, b | t_A, t_B) = \begin{cases} 0 & t_B^{\text{inf}} \leq t_A^{\text{inf}} \\ 1 & t_A^{\text{inf}} < t_B^{\text{inf}} \leq t_A^{\text{cull}} \\ e^{-b(t_B^{\text{inf}} - t_A^{\text{cull}})} & t_B^{\text{inf}} > t_A^{\text{cull}} \end{cases} \quad (2.2)$$

(e) Geography

We assume that farms are more prone to infect farms close by than far away, with the likelihood contribution of infecting a farm a certain distance $|x_A - x_B|$ kilometres away given by the best-fitting distance kernel taken from Boender *et al.* [15]:

$$L_{\text{geo}}(\delta_{AB}, r_0, \alpha | x_A, x_B) = \frac{1}{1 + (||x_A - x_B||/r_0)^\alpha}. \quad (2.3)$$

We performed the analysis with a wide range of different shapes for this kernel, and found the results to be robust to the specific choice of kernel (see the electronic supplementary material).

(f) Genetics

We assume the sequence sampled to be the most prevalent sequence on the whole farm. Additional cloning experiments performed by Bataille *et al.* [19] showed this to be very probable for four out of five farms tested, where the fifth was the first farm to be infected when the epidemic spread to the south of the country, thus having an exceptionally long infectious period. We assume there is a fixed number of nucleotides N that can mutate independently of each other, and any mutations get fixed during or shortly after infection. Since many mutations will drastically lower the fitness of the virus, we set N at one-third of the total number of nucleotides sequenced. We take p_{ts} and p_{tv} to be the expected number of transitions and transversions per infection, respectively. We assume there is a fixed probability p_{del} of a deletion occurring during any infection. Although such a deletion could decrease N , the length of the deletions is small enough to ignore this effect. Let the number of transitions and transversions needed to go from the sequence of A to that of B be denoted as d_{ts} and d_{tv} , respectively, and let $\mathbf{1}_{\text{del}}$ be an indicator function: 1 if a deletion occurred, 0 if it did not. We then have for the likelihood contribution:

$$L_{\text{gen}}(\delta_{AB}, \mathbf{p} | \text{RNA}_A, \text{RNA}_B) = \frac{(p_{\text{ts}}/N)^{d_{\text{ts}}} (p_{\text{tv}}/N)^{d_{\text{tv}}}}{(1 - (p_{\text{ts}}/N))^{d_{\text{ts}} - N} (1 - (p_{\text{tv}}/N))^{d_{\text{tv}} - N}} p_{\text{del}}^{\mathbf{1}_{\text{del}}} (1 - p_{\text{del}})^{1 - \mathbf{1}_{\text{del}}}. \quad (2.4)$$

(g) Likelihood with missing data

If all data were available for each farm, the likelihood of any transmission tree would be the product of the likelihoods of the links that it consists of. However, when data are missing for a certain farm, we have to incorporate information on the neighbouring farms in the tree to assess the likelihood. Figure 1c illustrates this concept. To make it precise, let T be the transmission tree to be evaluated. T consists of a set of links, one for each of the farms except the index case. Let i be a certain data type in the analysis, and let S_i be the largest partition of T into subsets, called subtrees, such that for each farm that misses data type i all links connected to that farm are in the same subtree. Then the likelihood of T is the product over all data types i of the product of the

likelihoods of each of the subtrees in S_i :

$$\begin{aligned} L(T, \mathbf{w} | D) &= L_t(T, b | \mathbf{t}_T) L_{\text{geo}}(T, r_0, \alpha | \mathbf{x}_T) L_{\text{gen}}(T, \mathbf{p} | \text{RNA}_T) \\ &= \prod_{S \in S_t} L_t(S, b | \mathbf{t}_S) \prod_{S \in S_{\text{geo}}} L_{\text{geo}}(S, r_0, \alpha | \mathbf{x}_S) \\ &\quad \prod_{S \in S_{\text{gen}}} L_{\text{gen}}(S, \mathbf{p} | \text{RNA}_S). \end{aligned} \quad (2.5)$$

The likelihood of a subtree is again the product of the likelihoods of its links, where for each of the farms with missing data, we sum over all possible data values. Note that if all data of a certain type are known for all farms, which for the epidemic of avian influenza is the case for both temporal and geographical data, the subtrees for that data type are just the individual links.

(h) Calculation

We construct a Monte Carlo Markov chain (MCMC) to sample from the space of all possible transmission trees and parameters $\mathbf{w} = (b, r_0, \alpha, p_{\text{ts}}, p_{\text{tv}}, p_{\text{del}})$, using flat priors for all parameters on the positive real numbers and for all transmission links. Subtrees containing unsequenced farms, i.e. farms without genetic data, are handled by summing over all data values consistent with the least amount of mutations needed to explain the subtree. Note that we do not have to sum over all unsequenced farms: first, genetic data are irrelevant for farms that infect no other farms (Y in figure 1). Second, farms that infect only one other farm (X in figure 1) can be handled by extending (2.4) to allow for a farm indirectly infecting another farm where there are x transmissions in the chain:

$$\begin{aligned} L_{\text{gen}}(\delta_{AB}, \mathbf{p} | \text{RNA}_A, \text{RNA}_B) &= \frac{x^{d_{\text{ts}}} (p_{\text{ts}}/N)^{d_{\text{ts}}} x^{d_{\text{tv}}} (p_{\text{tv}}/N)^{d_{\text{tv}}}}{(1 - (p_{\text{ts}}/N))^{d_{\text{ts}} - xN} (1 - (p_{\text{tv}}/N))^{d_{\text{tv}} - xN}} p_{\text{del}}^{\mathbf{1}_{\text{del}}} (1 - p_{\text{del}})^{1 - \mathbf{1}_{\text{del}}}. \end{aligned} \quad (2.6)$$

To correctly assess the likelihood of the parameters, we have to normalize the likelihood functions, i.e. divide the likelihood by the integral over the entire parameter range. Since the geographical locations of the farms are fixed, we normalize (2.3) for each farm by dividing by the total infecting potential for that farm, i.e. the sum over all other farms of the product of equations (2.2) and (2.3).

We take a burn-in of 500 000 iterations, checking for convergence, and then sample 10 000 times, at every 500th iteration. Averaging over the posterior density over the space of trees gives the probability for each possible infection event. We obtain a point estimate for parameters by taking the median of their posterior densities and construct a 95% credibility interval by taking 95% of the posterior probability mass.

For each of the sampled trees, we can estimate the infectiousness of each farm by dividing the number of infections caused by this farm by the length of its infectious period. Taking the average for each farm type gives us an estimate of relative infectiousness of different types of farms.

(i) Evaluation of the estimated tree

To evaluate how much information is contained in the geographical and genetic data, respectively, we exclude the geographical (2.3) or the genetic (2.4) likelihood from the full likelihood function (2.1), re-run the analyses, and see how much our results differ from the full picture obtained

Figure 2. Infection events with posterior probability greater than 0.5. The analysis was run separately three times using (a) temporal, genetic and geographical, (b) temporal and genetic or (c) temporal and geographical data. Red dots denote infected farms, yellow dots denote farms not infected in the epidemic. A higher arrows opacity corresponds to a higher estimated probability. The lack of arrows in (c) tells us geographical information alone is not enough to establish transmission links. Although genetic information yields quite accurate results, we obtain more certainty for many links when also incorporating the geographical data. Furthermore, in the combined analysis, we can correctly show there was only one introduction into the southern province of The Netherlands rather than multiple as suggested by the genetic data (indicated by the red arrows). A magnified version of the northern part of the outbreak is given in the electronic supplementary material, figure S1.

by using all data. Two important properties of the estimated tree to look at are correctness and resolution; how close the tree is to the actual transmission tree and how many links can be established at or above a certain probability level, respectively. Although correctness is the most important one, it is also the hardest to measure. Resolution tells us how well we can distinguish between potential source farms; if our assumptions on the disease dynamics are reasonable, an increased resolution should also be indicative of an increased correctness.

To test whether the method correctly handles cases with missing data, we analyse the position of the unsequenced farms in the estimated transmission tree.

3. RESULTS

From our analysis, we obtain an estimate of the transmission tree; for each pair of farms A and B, we get the probability that A infected B. A graphical representation of the estimated tree is given in figure 2a, which shows the high-probability transmissions on a map of the region. Figure 2b,c does the same for the trees obtained when excluding geographical and genetic data, respectively, from the analysis. The decrease in estimated transmissions relative to figure 2a shows that combining all data leads to an increase in resolution. An increase in correctness is also shown: the combined approach suggests only one introduction of the disease into the southern part of the country (figure 2a), while the analysis based on genetic data predicts multiple introductions (figure 2b). However, additional cloning experiments [19] have made clear that the disease was transmitted to the south of the country only once and then spread locally.

The resolution of the estimated trees is depicted in more detail in figure 3: it shows that more resolution can be obtained when using the genetic than when using the geographical data, but combining these gives the highest resolution. Furthermore, the figure makes clear that the 56 unsequenced farms cannot be placed in the tree with high confidence.

Figure 3. Resolution of the estimated trees. For each level of probability, the percentage of farms for which the farm that infected it can be estimated at or above this level is plotted. Estimates are based on temporal and (blue) genetic, (red) geographical or (black) both data. Dashed and dotted lines give the same information for sequenced and unsequenced farms, respectively. The graph for genetic data (blue) is much higher than that for the geographical data (red), which shows the former yields more resolution than the latter. This is confirmed by the fact that unsequenced farms are hard to place accurately in the transmission tree, resulting in a small surface under the dotted lines. However, combination of all data types results in the highest resolution, shown by the largest surface being under the black line.

There is a subtle point that can be made about the placement of unsequenced farms in the estimated tree. That is, if an unsequenced farm in reality infected another farm that we did sequence, we would observe a relatively large genetic distance between the farm that infected the unsequenced farm and the farm that was infected by it. Therefore, we would expect a sequenced farm that has

Figure 4. Scatterplot illustrating the position of unsequenced farms in the estimated tree. For each sequenced farm, we plot the minimum of the genetic distances to the sequenced farms infected earlier than this farm (horizontal axis) against the probability it was infected by an unsequenced farm (vertical axis). The small horizontal lines give the average probability per distance, the large horizontal line gives the fraction of unsequenced farms (equal to 0.23). The increasing trend shows that the placement of unsequenced farms in the estimated tree is more likely to be between farms whose genetic distance is relatively large. This is probably owing to large genetic distances being indicative of farms in the actual transmission tree not being sequenced.

a large genetic distance to all other farms to be actually infected by an unsequenced farm with higher probability, although in general, we cannot identify which of the unsequenced farms were responsible. Indeed, in the estimated tree, we see that sequenced farms that have a large genetic distance to other farms are more likely to be infected by an unsequenced farm (figure 4).

Table 1 gives our estimates of the parameters used to describe the dynamics of the disease. The estimated value of $b = 0.28$ shows that infectiousness drops by $(1 - \exp(-0.28)) \times 100\% \approx 26\%$ every day after culling. This low decay rate could be partly owing to errors in the estimation of the infection dates. The values of r_0 and α give the infection pressure exerted by a farm to farms a certain distance away, for example the pressure exerted at 1 km is 24 times higher than that exerted at 10 km. The expected number of point mutations per transmission is $p_{ts} + p_{tv} \approx 1.4$, reflecting the high genetic diversity found in this epidemic.

The transmission tree can be used to estimate the infectiousness of the different types of farms involved in the epidemic, and to look at the relation between farm size and infectiousness (figure 5). We see that hobby farms on average caused fewer, and turkey farms caused more infections per day than chicken farms. This last observation could be partly explained by the fact that turkey farms were among the first to be infected when the epidemic spread to the south of the country, where the infections remained undetected for some time. Although there is a relation between farm size and infectiousness, it takes a peculiar form. It seems that farms with fewer than 1000 animals (this includes all hobby farms) are hardly infectious when compared with

Table 1. Estimates of the parameters used, with 95% credibility intervals (CI).

parameter (units)	interpretation	estimated value (95% CI)
b (day ⁻¹)	rate of decline of infectiousness	0.28 (0.23, 0.34)
r_0 (km)	scale parameter of spatial kernel	2.4 (1.2, 3.7)
α	shape parameter of spatial kernel	2.3 (1.7, 2.8)
p_{ts}	average number of transitions	1.1 (0.88, 1.3)
p_{tv}	average number of transversions	0.32 (0.22, 0.43)
p_{del}	probability of deletion	0.069 (0.027, 0.13)

the larger farms, but above this boundary infectiousness does not increase with size.

4. DISCUSSION

We have estimated the transmission tree for an epidemic of avian influenza by combining genetic, geographical and temporal data using one likelihood function. We have shown the increased correctness and resolution obtained by using all data types in one analysis, illustrated that the approach correctly handles missing data, and estimated the infectiousness of farms of different types and sizes.

The results we obtained on the epidemic of avian influenza fit well into previous results. Bavinck *et al.* [16] found hobby farms to be less susceptible, but lacked the data to assess infectiousness, which we additionally showed to be lower than that of other farms. The dichotomous relation we found between farm size and infectiousness might be explained by large farms being detected and culled before a substantial number of animals could get infected, or by a mechanism of spread that is not very sensitive to the amount of infectious particles present in a farm. Boender *et al.* [15] used epidemiological data on all poultry farms in The Netherlands to estimate the parameters of the distance kernel (2.3). Even though they did not have any genetic information available, their estimates of r_0 and α , 1.9 (1.1, 2.9) and 2.1 (1.8, 2.4), are very close to our estimates, 2.4 (1.2, 3.7) and 2.3 (1.7, 2.8). The similarity of these estimates, based on different datasets, increases our confidence in both analyses. Further research on this H7N7 epidemic should focus on how the virus spread; knowledge on which farm infected which coupled with information on, for example, human movement and/or historical wind direction could answer long-standing questions on what mechanisms are responsible for the spread of avian influenza. Furthermore, it will be interesting to look at different predictors for infectiousness; although we lacked the data to do so, it might be very worthwhile to look at the impact of, for example, different contact structures or trading practices.

In modelling evolutionary drift, the usual assumption is that of a (relaxed) molecular clock, i.e. a constant rate at which new mutations fix in the population.

Figure 5. Estimated average infectiousness for (a) different types of farms and (b) number of animals on the farm, as measured by number of infections caused divided by the time period in days between infection and culling of the farm. All farms to the left of the dashed line in (b) are hobby farms by definition (these are farms with less than 300 animals). (a) We see hobby farms are less infectious than other types of farms, while turkey farms are more infectious than chicken farms. This can possibly be explained by the fact that many of the turkey farms in the epidemic were among the first to be infected when the disease spread to the southern part of the country, where control measures were not yet in place. (b) There is a correlation between total animals present on a farm and their infectiousness, however, the exact relationship remains unclear.

This assumption appears valid on long timescales; however, it is unclear whether this assumption holds on short timescales such as the duration of one avian influenza epidemic. We make the simplifying assumption that mutations only fix in the viral population during or shortly after infection, independent of time. The rationale behind our assumption is that the infection of a farm constitutes a population bottleneck for the virus. If infectious periods are short, these bottlenecks are exactly where the emergence of new dominant strains is to be expected, even though the particular mutations could have been present before the bottleneck. A more sophisticated evolutionary model that depends both on time and on infection events would be highly desirable, and constructing and fitting one will be a challenging improvement to the analysis in this paper.

When using different types of data, a natural question is how to 'weigh' these different types. We argue that the most natural approach to weighing is by constructing intuitively plausible likelihood functions for each of the data types and then combining these in a straightforward manner such as in our analysis. There is then a weighing of data types implicitly done by the choice of function. For example, the number of nucleotides that could mutate, N in equation (2.4), is not known exactly; increasing or decreasing this value puts more or less weight on the genetic data, respectively. Plausible numbers for N range from the total number of nucleotides seen to mutate (214) to the total number of nucleotides sampled (5354). We found the results described in this paper to be robust to changes in N for this range.

An alternative approach to estimating transmission trees using genetic and epidemiological data that have been suggested by Cottam *et al.* [10] is to assess the data types sequentially; first, using genetic information to exclude possible transmission trees, then assessing the likelihood of these using epidemiological data. In our terminology, this is equivalent to assigning each possible transmission tree a likelihood based solely on genetic

data, and discarding the trees with likelihood values below a certain level. Not only is this choice of level arbitrary, but since the remaining trees are evaluated only using epidemiological data, any further information in the likelihoods that were derived from the genetic data is lost. A third approach was presented by Jombart *et al.* [11], who nicely contrasted the construction of transmission trees with the construction of phylogenetic trees, making the link to the field of phylogeography. They, however, only look at epidemiological data when genetic sequences for multiple cases are identical, disregarding these data otherwise. Furthermore, the method lacks a way to handle missing data.

The general framework presented in this paper is applicable in a wide range of settings. Although we have looked at infections between farms, the same techniques are also applicable when the infections are between individual host organisms, such as humans, animals or plants. Different types of epidemiological data such as day of symptom onset, age or area code could be handled by constructing an appropriate likelihood function, for example, as in Wallinga & Teunis [4] and Mossong *et al.* [22]. Furthermore, the ability to handle missing data makes the method suitable for use in many practical settings. Since the information contained in genetic data increases with genetic diversity, the method is probably more suitable to analyse epidemics of rapidly mutating pathogens such as viruses; however, there are no theoretical objections that limit its application to epidemics of other types of pathogens, which could be worthwhile if much epidemiological data are available.

The framework presented in this paper reflects our belief that all data gathered on an epidemic can provide clues to what actually happened. How much information each data type holds is usually hard to determine, but the better we can extract this information, the clearer our picture of the epidemic becomes. Ultimately, techniques such as those presented here coupled with our increasing capacity to build up large (genetic) datasets will allow for

the accurate and correct estimation of transmission trees in a variety of settings, increasing our understanding of disease dynamics. The real challenge then lies in not only understanding these dynamics, but in also putting this knowledge to use in our efforts to combat diseases.

We would like to thank Annemarie Bouma, Gert-Jan Boender, Thomas Hagenaars and Michiel van Boven for critically reading an earlier version of this manuscript. We would like to thank the Associate Editor and two anonymous referees for useful comments.

REFERENCES

- Keeling, M. J., Woolhouse, M. E., May, R. M., Davies, G. & Grenfell, B. T. 2003 Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136–142. (doi:10.1038/nature01343)
- Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. 2001 Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature* **413**, 542–548. (doi:10.1038/35097116)
- Heijne, J. C., Teunis, P., Morroy, G., Wijkman, C., Oostveen, S., Duizer, E., Kretzschmar, M. & Wallinga, J. 2009 Enhanced hygiene measures and norovirus transmission during an outbreak. *Emerg. Infect. Dis.* **15**, 24–30. (doi:10.3201/1501.080299)
- Wallinga, J. & Teunis, P. 2004 Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516. (doi:10.1093/aje/kwh255)
- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E. & Getz, W. M. 2005 Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359. (doi:10.1038/nature04153)
- Spada, E., Sagliocca, L., Sourdis, J., Garbuglia, A. R., Poggi, V., De Fusco, C. & Mele, A. 2004 Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *J. Clin. Microbiol.* **42**, 4230–4236. (doi:10.1128/JCM.42.9.4230-4236.2004)
- Zhang, L., Diaz, R. S., Ho, D. D., Mosley, J. W., Busch, M. P. & Mayer, A. 1997 Host-specific driving force in human immunodeficiency virus type 1 evolution *in vivo*. *J. Virol.* **71**, 2555–2561.
- Leitner, T. & Albert, J. 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl Acad. Sci. USA* **96**, 10 752–10 757. (doi:10.1073/pnas.96.19.10752)
- Pybus, O. G. & Rambaut, A. 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550. (doi:10.1038/nrg2583)
- Cottam, E. M., Thebaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., King, D. P. & Haydon, D. T. 2008 Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B* **275**, 887–895. (doi:10.1098/rspb.2007.1442)
- Jombart, T., Eggo, R. M., Dodd, P. J. & Balloux, F. 2011 Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* **106**, 383–390. (doi:10.1038/hdy.2010.78)
- Stegeman, A., Bouma, A., Elbers, A. R., de Jong, M. C., Nodelijk, G., de Klerk, F., Koch, G. & van Boven, M. 2004 Avian influenza A virus (H7N7) epidemic in The Netherlands in 2003: course of the epidemic and effectiveness of control measures. *J. Infect. Dis.* **190**, 2088–2095. (doi:10.1086/425583)
- Fouchier, R. A. *et al.* 2004 Avian influenza A virus (H7N7) associated with human conjunctivitis and a fatal case of acute respiratory distress syndrome. *Proc. Natl Acad. Sci. USA* **101**, 1356–1361. (doi:10.1073/pnas.0308352100)
- Koopmans, M. *et al.* 2004 Transmission of H7N7 avian influenza A virus to human beings during a large outbreak in commercial poultry farms in the Netherlands. *Lancet* **363**, 587–593. (doi:10.1016/S0140-6736(04)15589-X)
- Boender, G. J., Hagenaars, T. J., Bouma, A., Nodelijk, G., Elbers, A. R., de Jong, M. C. & van Boven, M. 2007 Risk maps for the spread of highly pathogenic avian influenza in poultry. *PLoS Comput. Biol.* **3**, e71. (doi:10.1371/journal.pcbi.0030071)
- Bavinck, V., Bouma, A., van Boven, M., Bos, M. E., Stassen, E. & Stegeman, J. A. 2009 The role of backyard poultry flocks in the epidemic of highly pathogenic avian influenza virus (H7N7) in the Netherlands in 2003. *Prev. Vet. Med.* **88**, 247–254. (doi:10.1016/j.prevetmed.2008.10.007)
- Bessell, P. R., Shaw, D. J., Savill, N. J. & Woolhouse, M. E. 2010 Estimating risk factors for farm-level transmission of disease: foot and mouth disease during the 2001 epidemic in Great Britain. *Epidemics* **2**, 109–115. (doi:10.1016/j.epidem.2010.06.002)
- Bos, M. E., van Boven, M., Nielen, M., Bouma, A., Elbers, A. R., Nodelijk, G., Koch, G., Stegeman, A. & de Jong, M. C. 2007 Estimating the day of highly pathogenic avian influenza (H7N7) virus introduction into a poultry flock based on mortality data. *Vet. Res.* **38**, 493–504. (doi:10.1051/vetres:2007008)
- Bataille, A., van de Meer, F., Koch, G. & Stegeman, A. 2011 Evolutionary analysis of inter-farm transmission dynamics in a highly pathogenic avian influenza outbreak. *PLoS Pathog.*
- de Jong, M. C., Stegeman, A., van der Goot, J. & Koch, G. 2009 Intra- and interspecies transmission of H7N7 highly pathogenic avian influenza virus during the avian influenza epidemic in The Netherlands in 2003. *Rev. Sci. Tech.* **28**, 333–340.
- Spekreijse, D., Bouma, A., Stegeman, J. A., Koch, G. & de Jong, M. C. M. 2010 The effect of inoculation dose of a highly pathogenic avian influenza virus strain H5N1 on the infectiousness of chickens. *Vet. Microbiol.* **147**, 59–66. (doi:10.1016/j.vetmic.2010.06.012)
- Mossong, J. *et al.* 2008 Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74. (doi:10.1371/journal.pmed.0050074)