Robust Reconstruction and Analysis of Outbreak Data: Influenza A(H1N1)v Transmission in a School-based Population

Niel Hens*, Laurence Calatayud, Satu Kurkela, Teele Tamme, and Jacco Wallinga

* Correspondence to Dr. Niel Hens, Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Agoralaan 1, Building D, 3590 Diepenbeek, Belgium (e-mail: niel.hens@uhasselt.be).

Initially submitted April 6, 2011; accepted for publication May 5, 2012.

The rapid spread of the new influenza virus A(H1N1)v in young age groups in 2009 has been partly attributed to a high transmission intensity in schools. However, detailed characterization of the spread of influenza in school populations has been difficult to obtain, simply because it is very hard to identify who infected whom in a large outbreak. Data collected in large outbreak investigations typically miss many transmission events, and some reported transmission events will be incorrect. Here the authors present robust likelihood-based methods that can be used to analyze outbreak data while explicitly accounting for both missing data and erroneous data. They apply this method to a school-based outbreak of pandemic influenza A(H1N1)v that occurred in London, United Kingdom, in April 2009. The authors show that the generation interval in this school-based population was 2.20 days and that the reproduction number declined coincident with school closure, from 1.33 secondary cases per primary case to 0.43 secondary cases per primary case. These results provide quantitative evidence for the change in influenza transmission that is to be expected from school closure.

basic reproduction number; communicable diseases; disease outbreaks; disease transmission, infectious; epidemiologic methods; influenza A virus, H1N1 subtype; missing data

Abbreviations: CI, confidence interval; PEM, prior-based expectation maximization.

The advent of the in...
The change in reproduction number allows us to infer the impact of control measures on transmission intensity.

However, the data on an outbreak are typically incomplete and sometimes inaccurate: The transmission of infection is almost always an unobservable event, and the infector often remains unknown. The mean generation interval can still be calculated as the average duration between the symptom onset dates for known pairs of infector and infected (a quantity that is sometimes referred to as the “clinical onset serial interval” (9)). This estimate can suffer from bias in the detection of transmission events: When long generation intervals are easily missed, the estimate will tend to be shorter. For these reasons, most estimates of transmission characteristics are based on relatively small but well-documented outbreaks, where the methods are tailored to the data at hand, and most available estimators for the generation interval and reproduction numbers do not account for missing data and are not robust to errors in the underlying data. To the best of our knowledge, no available estimators for the generation interval and reproduction numbers use the additional information on the relation between cases and their contacts that is collected by outbreak investigators to assist in tracing the infector (1).

Here we show that it is possible to reconstruct transmission trees from data routinely collected in outbreak investigations and to estimate transmission characteristics such as the generation interval and the reproduction number. The trick we rely on is to explicitly consider the transmission events as observable events—even though they will be missing in most instances—and to use modern missing-data techniques to deal with the missing values, while explicitly allowing for the possibility that these observations are incorrect. In dealing with the missing data, we make use of the information we do have about the case, such as the list of contacts, and make use of the likelihood that the time between symptom onset in a case and symptom onset of the case’s infector is consistent with the few observed transmission pairs. We illustrate such an approach using data collected during an investigation of one of the first outbreaks of influenza A(H1N1)v virus in the United Kingdom. In this analysis, we are particularly interested in the generation interval in a school-based population, and we want to know how much the reproduction number among the school population is lowered when the school closes.

DATA

Early on in the spread of the pandemic influenza A(H1N1)v virus, an imported case was detected in a school with 1,177 pupils in London, United Kingdom (1). The index case was a pupil who had just returned from travel abroad and developed the first symptoms of infection on April 25, 2009 (day 0 of the outbreak). One of the Health Protection Agency’s local health protection units in London was alerted to potential cases of pandemic A(H1N1)v on April 29, 2009 (day 4 of the outbreak). As further cases were identified, the school was closed from May 2 to May 10, 2009 (days 7–15 of the outbreak). Mass antiviral prophylaxis was implemented by distributing antiviral drugs (oseltamivir) to pupils and staff on May 4 and May 5, 2009 (days 9 and 10 of the outbreak). Active searching and follow-up of cases and their close contacts was undertaken by the local health protection unit, with support from other colleagues in the Health Protection Agency.

The study population included any individual who attended the school as a pupil or worked there as a staff member between April 15 and May 15, 2009, as well as their close contacts. In total, 33 cases were identified within this study population. For these cases, information on the time of symptom onset was collected, and their contacts were identified. A case was defined as an individual in the study population who showed symptoms of influenza-like illness and tested positive for influenza virus A(H1N1)v by means of viral culture, polymerase chain reaction, or serologic diagnosis. A contact was defined according to United Kingdom guidelines as: any individual who lived in the same household as the case; any individual who provided informal care to the case, coming within 1 m of the case; any individual who had been exposed to a case at a distance smaller than 1 m, with continuous exposure over 1 hour; or any health-care or social worker who provided care or examined a symptomatic case.

The case reports included the following information (1):

- A unique case identifier, i. We rank cases by symptom onset date, with case $i = 1$ for the index case and case $i = n$ for the last case. Without loss of generality, we assign a specific chronologic order for those cases with equal symptom onset dates.
- Time of symptom onset for case $i$, $t_i$. We count time in days, starting counting at day $t = 0$ when the index case has onset of symptoms.
- The case identifier of the infector of case $i$, $v(i)$. For many cases, this information is missing.
- The generation interval for case $i$, as the duration between symptom onset of case $i$ and symptom onset of its infector $v(i)$, $t_i - t_{v(i)}$. If information on the infector is missing, data on the generation interval are missing as well.
- The case identifiers of the contacts of case $i$, $w(i)$.

DATA ANALYSIS

Likelihood of transmission pairs, ignoring missing data

We assume that observed generation intervals are positive random variables $t_2 - t_{v(2)}, \ldots, t_n - t_{v(n)}$ that are well described by a generation interval distribution $g(t_i - t_{v(i)} | \theta)$, where the parameters $\theta$ specify the location and shape of that distribution.

If we had complete case records, we would know for each case $i$ the infector, $j = v(i)$, and we could calculate the generation interval as $t_i - t_j$. We can use the information on the infectors $v$ to make an indicator variable $p_{ij}(v)$ that takes the value 1 if case $i$ is infected by case $j$ and 0 otherwise. We would write the log-likelihood function for the
generation intervals as

\[ \ell(\theta|t, v) = \sum_{i=2}^{n} \sum_{j=1}^{n} p_{ij}(v) \log g(t_i - t_j|\theta). \]  

However, we have incomplete case reports, and for many cases the infector \( v(i) \) is missing. A naive approach is to simply find the best-fitting generation interval distribution for only observed generation intervals, which requires an implicit assumption that generation intervals are missing completely at random. Such an assumption is untenable whenever it is plausible that infectors are more easily missed when the generation interval is either unusually short or unusually long. Moreover, considering observed generation intervals only leads to larger uncertainty because of the sample size reduction.

**Likelihood of possible transmission pairs, dealing with missing data**

A better, more robust way of dealing with missing data is to find the best-fitting generation interval for both observed and missing generation intervals. This can be done via the expected log-likelihood of generation intervals, where expectation is over the possible values that missing generation intervals can take. To calculate these possible values of missing generation intervals, we use the probability \( p_{ij}(v, w) \) that case \( j \) is the infector of case \( i \), given the duration between symptom onset of case \( i \) and case \( j \), given the information on the known possible infectors \( v \) and the known contacts \( w \).

The total log-likelihood of complete data (that is, both observed and missing data) is then the total of all contributions of all cases \( i = 2, \ldots, n \) to the log-likelihood:

\[ E\{\ell(\theta|t, v, w)\} = \sum_{i=2}^{n} \sum_{j=1}^{n} p_{ij}(v, w) \log g(t_i - t_j|\theta). \]  

Maximizing this expected log-likelihood \( E\{\ell(\theta|t)\} \) gives the maximum-likelihood estimates \( \hat{\theta} \) for the parameters that specify the generation interval distribution.

The probability that case \( i \) is infected by case \( j \), \( p_{ij} \), can be calculated as the probability of observing the duration between the symptom onsets in cases \( i \) and \( j \) given infectious contact, \( g(t_i - t_{v(i)}|\theta) \), times the probability that there was a potentially infectious contact between \( i \) and \( j \), \( \pi_{ij} \), normalized by the probability of \( i \)'s being infected by any other case \( k \):

\[ p_{ij}(v, w, \hat{\theta}) = \frac{g(t_i - t_j|\hat{\theta}) \times \pi_{ij}(v, w)}{\sum_{k \neq i} g(t_i - t_k|\hat{\theta}) \times \pi_{ik}(v, w)}. \]  

The contact information that is collected during the outbreak informs us on \( \pi_{ij}(v, w) \), the probability of a potentially infectious contact between cases \( i \) and \( j \): If the outbreak investigation reveals that \( i \) and \( j \) form a likely transmission pair, that is, if \( v(i) = j \), there is only 1 possible infector, and hence the probability of a potentially infectious contact is \( \pi_{ij}(v, w) = 1 \). If the outbreak investigation reveals that \( j \) is one out of \( m \) contacts of \( i \), there are \( m \) possible infectors, and the probability of a potentially infectious contact is \( \pi_{ij}(v, w) = 1/m \). If the outbreak investigation reveals that \( i \) is not an index case and did not contact other cases, all \( i - 1 \) cases that developed symptoms before case \( i \) are possible infectors, and the probability of a potentially infectious contact is \( \pi_{ij}(v, w) = 1/(i - 1) \).

The expected log-likelihood function can be maximized by first evaluating the probability of potentially infectious contacts given the contact information from outbreak investigation data (P-step), then evaluating the probabilities of transmission of infection (E-step) given the current best estimate of the generation interval parameters \( \theta \), and then finding the value of the generation interval parameters \( \theta \) that maximize the expected likelihood given the probabilities of transmission of infection (M-step). The E and M steps are repeated until the results converge to the maximum expected log-likelihood value (10). This iterative procedure is referred to as the prior-based expectation maximization (PEM) algorithm.

**Identification of unlikely transmission pairs**

It is very well possible that some observed generation intervals are incorrect, and these incorrect generation intervals may have a large impact on the estimated mean generation interval. We can identify the impact of 1 single observed generation interval on the inference process, by changing the data on the infector \( v(i) \) to a missing value, and consider the change in the expected log-likelihood of the parameter \( \theta \). This is a so-called “global influence measure,” and we can write this more formally as \( G_I = E\{\ell(\theta|v(i) = \text{missing})\} - E\{\ell(\theta)\} \). Cases with a large value for the global influence measure should be regarded with suspicion and should be checked for transcription errors or other explanations of apparent anomalies. In the absence of any explanation, they may possibly be left out of the analysis.

**Estimation of the mean generation interval**

We determine the maximum-likelihood value \( \hat{\theta} \) for the parameters of the generation interval distribution and calculate the mean generation interval as

\[ T = E\{g(t_i - t_j|\hat{\theta})\}. \]  

This expected duration between symptom onset dates from infected back to infector is equal to the expected duration between infection times from infected back to infector, provided that the incubation times of successive cases are independent (11).

We can postulate any statistical distribution for the generation interval distribution \( g \) and assess its adequacy in describing the data, whenever we can ensure that the cases \( i \) and \( j \) are located on the same transmission tree, and thus avoid the causal possibility that case \( i \) is infected by itself or by one of its secondary cases. Here we choose to describe the generation interval distribution for
Influenza by a gamma distribution. The distribution has 2 parameters: a shape parameter \(a\) and a scale parameter \(b\), such that the mean generation interval is \(T = ab\) and the variance of generation intervals is \(\sigma^2 = ab^2\). The distribution is nonnegative, and this guarantees that all causal impossible infection events are assigned a zero likelihood. We obtained similar results when analyzing the data using other 2-parameter distributions such as the Weibull or log-normal distribution.

Reconstruction of possible transmission trees and the time course of reproduction numbers

We have for each case \(i\) the probabilities that case \(i\) is infected by case \(j\): \(\{p_{ij}, \ldots, p_{mj}\}\). These probabilities specify a multinomial distribution, and by sampling 1 draw from such a distribution, we can simulate a possible infecter of case \(i\) and use this as a “bootstrapped” infecter \(v(i)^{\ast}\). Doing this for all cases (except the index case) will give us a tree in which all cases are connected and the index case is the root of the transmission tree. For each bootstrapped transmission tree, we count for each case the number of secondary cases and use this as a case-reproduction number. We estimate the reproduction number on day \(t\), \(R_t\), as the average reproduction number of all cases who have symptom onset on day \(t\), as in the paper by Wallinga and Teunis (7):

\[
R_t = \sum_j \sum_{i=2} p_{ij}(v, w, \theta).
\]

Testing the estimation procedure with simulated data

We checked the accuracy of the estimation procedure using the PEM algorithm in a simulation study against alternative methods. We found that the proposed PEM algorithm outperformed alternatives in terms of the mean averaged squared error of the estimate, even when missingness of data on possible infecters depended on the length of the actual generation interval (see the Web Appendix, which appears on the Journal’s website (http://aje.oxfordjournals.org/), for details). We also observed that the algorithm is very fast, requiring only a few seconds on a laptop computer with a 2.20-GHz processor and 8 GB of RAM for the data set of 33 cases (see Web Appendix for details).

RESULTS

Figure 1 shows the symptom onset dates of the 33 cases in the 2009 influenza outbreak at the school in London. For 7 cases, the outbreak investigators observed a unique likely infecter (indicated by a drawn line). The estimated mean of the 7 observed generation intervals was 2.43 days (95% confidence interval (CI): 1.41, 4.17; Figure 2).

We reconstructed plausible transmission trees for the outbreak (one such transmission tree is depicted in Figure 3). The estimated mean generation interval was 2.20 days (95% CI: 1.32, 3.68); these estimates were based on 7 observed generation intervals and were consistent with 25 unobserved generation intervals (Figure 4A). Application of the global influence measure revealed that the generation intervals for 2 likely transmission pairs (C39–C52 and C38–C53) were highly influential. Removing the source cases for these suspected pairs and treating these generation intervals as missing resulted in a mean generation interval of 2.22 days. The impact of suspected observation on the estimate was small and showed that the estimated mean generation interval of 2.20 days was robust to erroneous observations.

To assess the impact of school closure on transmission, we estimated the time course of reproduction numbers. The case reproduction numbers decreased as time progressed (Figure 4B). The reproduction number before detection of the outbreak was 2.51 (95% CI: 2.11, 3.00; estimate based on cases with symptom onset days 0–4); the reproduction number after detection of the outbreak but before school closure was 1.33 (95% CI: 1.11, 1.56; estimate based on cases with symptom onset days 5–7); and the reproduction number after detection of the outbreak and after school closure was 0.43 (95% CI: 0.35, 0.52; estimate based on cases with symptom onset days 8–12).

DISCUSSION

We reconstructed and analyzed the transmission characteristics of influenza A(H1N1)v in a school-based population in London during the 2009 outbreak. We used data collected during the outbreak investigation that were, by nature, incomplete and may have contained errors. We employed the data on the contacts of cases, which are normally excluded from any analysis, to deal with missing data and trace possible errors in the data. This allowed us to infer a plausible transmission tree and the mean generation interval. It also allowed us to infer the change in the reproduction number when the outbreak was detected and the school was closed.

Using only the completely observed transmission pairs of primary and secondary cases, without correcting for missing data, we would have estimated a mean generation interval of 2.4 days. This is close to the estimate of 2.7 days for an influenza outbreak at a New York City school that was also obtained using complete transmission pairs only (4). If we had accounted for the missing data without using auxiliary information, we would have estimated a mean generation interval of 2.1 days. This is closer to the estimate of 1.5 days that was reported for an outbreak at a school in Pennsylvania, which was obtained using a Bayesian approach that accounted for missing data (3). As is shown in the Web Appendix, such an estimate can be improved by invoking the auxiliary information on the contacts of the cases. This results in an estimate of the mean generation interval of 2.2 days.

The influenza reproduction number in the school-based population in London decreased significantly during the outbreak, by a factor of 0.3, from \(R_0 = 1.33\) when the outbreak was discovered to \(R_t = 0.43\) after the school was closed. In the school-based population in Pennsylvania, the reproduction number also decreased by a factor of 0.3, but...
this change was not statistically significant (3). In the general population of Hong Kong, the reproduction number decreased by a factor of 0.25 when schools closed (6). It is important to realize that we are analyzing a small outbreak (33 cases) and that there are a number of factors other than school closure per se that could have caused a decrease in the reproduction number. For example, in a large outbreak, the expected case reproduction numbers of the first cases are higher than those of subsequent cases; the reproduction numbers will decrease as the study population is depleted of susceptible individuals; and the estimated reproduction numbers will be biased toward lower values when infection

Figure 1. Reported transmission events during an outbreak of influenza A(H1N1)v at a school in the United Kingdom, April 2009. Cases (C) are indicated by boxes and are ranked horizontally by date of symptom onset. Reported transmission events between cases are indicated by solid arrows between boxes. For all cases without a reported infector, the contacts with other cases are indicated by dotted lines.

Figure 2. Relative frequency of generation intervals and observed number of secondary cases per primary case during an outbreak of influenza A(H1N1)v at a school in the United Kingdom, April 2009. A) Relative frequency of generation intervals for the 7 observed transmission events (bars) and the best fit of a gamma distribution to these transmission events (solid line). These data summaries do not account for the 25 unobserved transmission events. B) Observed number of secondary cases per primary case. Because only a small proportion of all transmission events is reported, this yields a lower bound for the reproduction number. The sizes of the dots are proportional to the numbers of cases.
spreads outside of the study population and secondary cases are missed. We think it is unlikely that these other factors could account for the observed decrease in the reproduction number, but since we cannot completely exclude them, our study suggests that the data are consistent with a decrease in the reproduction number of, at most, 0.3 in the school population upon closure of the school.

Outbreak investigators must walk a fine line between overinterpreting and ignoring the available data, which may be imperfect and incomplete. To provide a proper sense of the evidence contained in these data, it is important that a statistical approach used to analyze outbreak data is robust in the sense that it can deal with data that have partially incomplete and partially incorrect information on the possible

Figure 3. A plausible transmission tree with reconstructed transmission events for an outbreak of influenza A(H1N1)v at a school in the United Kingdom, April 2009. Cases (C) are indicated by boxes and are ranked horizontally by date of symptom onset. The reconstructed transmission events between cases are indicated by solid arrows. Transmission events were reconstructed using a missing-data method (the prior-based expectation maximization algorithm) that accounts for the likelihood of the duration of the generation interval and reported contact between cases.

Figure 4. Reconstructed transmission of influenza virus during an outbreak of influenza A(H1N1)v at a school in the United Kingdom, April 2009. A) Relative frequency of the generation intervals, based on the 7 observed transmission events while accounting for the 25 unobserved transmission events; B) time course of the average case reproduction number, \( R_t \). Bars, 95% nonparametric bootstrap percentile confidence interval.
infectors of the cases. Here we controlled for potential bias that may arise in a missing-data analysis by invoking the auxiliary information on contacts of the cases that was collected by the outbreak investigators during contact investigations. The contact information and the time of symptom onset, taken together, are used to find the most likely infector of each case. This allows for “nonrandom dropout” (12) of generation intervals and requires very mild assumptions about why information on the infector is missing. Such a missing-data approach can be used to construct a global influence measure (or a case-deletion measure) that elicits suspicious observations where the observed infector for a case might be incorrect. This information can be of direct help to the outbreak investigators.

The approach to analyzing outbreak data as proposed here can be extended to other infections that cause outbreaks, such as norovirus, or other emerging infections like ebola or pneumonic plague. Only a few assumptions about the infectious disease outbreak should be met. The analysis assumes that all cases in the outbreak, except the index case, are infected by another case within the study population. This assumption is plausible for any rare, emerging infection that spreads from human to human. Furthermore, the analysis assumes that all infectious cases are observed. This assumption is met when, as is the case for influenza, infectiousness is correlated with the severity of symptoms. Finally, the analysis assumes that the generation interval distribution remains stationary. Because there is no a priori justification for such an assumption from epidemiologic theory (11, 13–16), it is important to inspect the reconstructed durations of the individual generation intervals for deviations from stationarity. Here we did not find statistical evidence for such a deviation.

The availability of a fast and robust method with which to analyze data that are routinely collected during outbreak investigations will open up the possibility of addressing several important questions. First, it is no longer necessary to restrict analyses to small outbreaks where the outbreak data are complete, and outbreak analyses can be performed for larger outbreaks with data sets that will contain many missing values and errors. Second, there is a need to characterize the heterogeneity in the generation interval distribution and case reproduction numbers for transmission events in different settings and between different groups. Ongoing work suggests that for influenza, there is indeed a difference in the generation interval by setting (household, school, work (3)). Third, the ability to include prior information on the likelihood of transmission of infection between any 2 cases in the PEM algorithm allows us to incorporate information on molecular typing of the pathogens of these 2 cases. We can indicate in the analysis that cases with a similar pathogen genotype could have infected each other, whereas cases with very different pathogen genotypes could not have infected each other. This opens up perspectives for studying the transmission characteristics of infections that can be asymptomatic and where the duration of the generation interval distribution can be very long—for example, for infections such as hepatitis B or tuberculosis.

The analysis of influenza in a school-based population that is presented here provides us with better estimates of the generation interval and with an estimate of the reduction in transmission upon school closure. Perhaps more importantly, it also shows how we can learn from data that are routinely collected in outbreak investigations. Usually, only the data on cases’ time of symptom onset (the epidemic curve) are published, and the data on likely infectors and data on possible contacts of the cases are discarded. We show how this information proves highly valuable in extracting transmission characteristics of the infection. We would strongly advise against discarding information on infectors and contacts, and we encourage making this information available for careful statistical analysis, such that the experience of each outbreak investigation can be used to improve infection control.

ACKNOWLEDGMENTS

Author affiliations: Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium (Niel Hens, Telee Tamme); Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine and Infectious Disease Institute (World Health Organization Collaborating Centre), University of Antwerp, Antwerp, Belgium (Niel Hens); Health Protection Agency, London, United Kingdom (Laurence Calatayud, Satu Kurkela); European Programme for Intervention Epidemic Training, European Centre for Disease Control, Stockholm, Sweden (Laurence Calatayud); European Public Health Microbiology Training Programme, European Centre for Disease Control, Stockholm, Sweden (Satu Kurkela); Centre for Infectious Disease Control Netherlands, National Institute for Public Health and the Environment, Bilthoven, the Netherlands (Jacco Wallinga); and Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, the Netherlands (Jacco Wallinga).

N. H. was supported by SIMID (Simulation Models of Infectious Disease Transmission and Control Processes), a strategic basic research project funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders, Belgium (project 060081), and by the Belgian government’s IAP (Interuniversity Attraction Poles) research network (Belgian Science Policy Office) (grant P6/03). N. H. also received support from the University of Antwerp Scientific Chair in Evidence Based Vaccinology. J. W. received research funding from the European Union FP7 FluModCont Project (grant 20160). S. K. was supported by the European Public Health Microbiology Training Programme, which was developed and endorsed by the establishment of a collaborative network of European laboratories for outbreak assistance and support, coordinated by the European Network for Diagnostics of “Imported” Viral Diseases and the European Centre for Disease Prevention and Control.

The authors thank Drs. Marc Baguelin, Helen Maguire, Richard Pebody, Albert Jan van Hoek, and Peter J. White for their valuable contributions. The authors are also grateful to staff at the Health Protection Agency’s South East London Health Protection Unit and the South London Specialist Virology Centre, Health Protection Agency London Regional.
Laboratory, King’s College Hospital NHS Foundation Trust, for collecting the data and making them available.

Portions of this paper were presented at a workshop (“Design and Analysis of Infectious Disease Studies”) held at the Mathematisches Forschungsinstitut Oberwolfach, Oberwolfach, Germany, November 1–7, 2009.

Conflict of interest: none declared.

REFERENCES