Patterns and scales in infectious disease surveillance data: an exploratory data analysis approach

Soetens, L.C.

5

Real-time estimation of epidemiologic parameters from contact tracing data

The contents of this chapter have been published in *Epidemiology*:
Real-time estimation of epidemiologic parameters from contact tracing data during an emerging infectious disease outbreak
Loes Soetens, Don Klinkenberg, Corien Swaan, Susan Hahné, Jacco Wallinga

Epidemiology, March 2018, 29(2):230-236

Abstract

Contact tracing can provide accurate information on relevant parameters of an ongoing emerging infectious disease outbreak. This is crucial to investigators seeking to control such an outbreak. However, crude contact tracing data are difficult to interpret and methods for analyzing these data are scarce. We present a method to estimate and visualize key outbreak parameters from contact tracing information in real time by taking into account data censoring.

Exposure type-specific attack rates and the reproduction number *R(t)* are estimated from contact tracing data by using maximum likelihood estimation to account for censored data. The attack rates reflect, in the context of contact tracing, the specificity of the contact definition; a higher value indicates relatively efficient contact tracing. The evolution of *R(t)* over time provides information regarding the effectiveness of interventions. To allow a real-time overview of the outbreak, the attack rates and the evolution of *R(t)* over time are visualized together with the case–contact network and epicurve. We applied the method to a well-documented smallpox outbreak in the Netherlands, to demonstrate the added value. Our method facilitates the analysis of contact tracing information by quickly turning it into accessible information, helping outbreak investigators to make real-time decisions to more effectively and efficiently control infectious disease outbreaks.

Introduction

In an ongoing emerging infectious disease outbreak, contact tracing can provide accurate real-time information on relevant parameters, which is needed by public health professionals to control an outbreak [1-4]. Especially in an emerging infectious disease outbreak, active daily (or even hourly) monitoring of contacts provides a wealth of information. First, data on how many contacted individuals (hereafter: contacts) remain under surveillance can indicate what control efforts are possibly required in the short term. Second, data on exposure type-specific attack rates (ARs) can help in targeting control efforts to contacts at high risk. We defined the exposure type-specific AR as the proportion of infected individuals with a certain exposure type among the total number of traced individuals with that exposure type. Third, estimation of the effective reproduction number *R(t)*, defined as the expected number of secondary cases per primary case infected at time *t* [5], allows evaluation of infection transmissibility and the effect of current control measures.

However, real-time use of contact tracing data by public health professionals is limited, despite publication of several methods for inferring key parameters from outbreak data [6-13]. One problem is that contact tracing data contains a lot of information on confirmed cases and their contacts through various exposure routes at different time-points or intervals. Using all that information requires an overview. In addition, many methods focus on just a single parameter [7-10], often the effective reproduction number. To obtain all key outbreak parameters, a public health professional has to search through a wide range of methods, which is quite cumbersome and time-consuming. A third issue is that real-time analysis should account for the fact that the data are typically censored: at the time of analysis, the outcome (whether a contact is infected or not) is unknown for many exposed persons [9]. Censoring has a consequence for accurate estimation of outbreak parameters, such as ARs, which may be underestimated if traced contacts later appear to be cases. As far as we know, only few studies take such censoring into account [9, 10]. A final issue is that published methods require specific software that is either hard to obtain or requires advanced programming skills [8, 12, 13].

To improve public health decision-making during an emerging infectious disease outbreak, we developed a tool that uses contact tracing information to estimate outbreak parameters in real time while correcting for right censoring of the data (i.e. taking into account that contacts in follow-up may become cases in the near future in the real-time estimation of outbreak parameters). Our tool provides a comprehensive visualization of the outbreak, including the course of the outbreak and contacts under surveillance. The tool is easy to use by public health professionals as it is available in a user-friendly web interface. The tool is specifically aimed at analyzing contact tracing data of emerging infectious disease outbreaks, as contact tracing is most relevant during such outbreaks where there is the

greatest need for such knowledge about the ongoing outbreak. Our tool is aimed at serious infectious disease outbreaks, for which active daily contact tracing is in place, whereby all cases are identified. In this paper, we demonstrate this tool, using contact tracing data from a well-documented smallpox outbreak in the Netherlands in 1951.

Materials and methods

Data

Data for the tool should be entered in a spreadsheet, with a row for each case or contact and the following information in the columns:

- unique identifier (ID)
- gender
- infection status: is it a case or a contact?
- ID of source
- exposure type
- exposure date (first day)
- exposure duration
- date of disease onset

In this paper, we use the word “contact” to mean an individual who has come into contact with a case and has not yet been shown to be a case himself and “exposure” to mean the event of contact between the two individuals. Additional information that should be obtained is the incubation period distribution (mean and standard deviation) and the generation interval distribution (mean and standard deviation). The generation interval refers to the time between symptom onset of successive cases in a chain of transmission. The data saved as a csv-file can be uploaded into a user-friendly web interface for easy use of the tool (https://github.com/lsoetens/Contactviz) or loaded via the R code provided at the same place, using the R statistical software.

For illustration of the tool, we used data from a smallpox outbreak in Tilburg, The Netherlands, in 1951 [14, 15]. In this outbreak, 52 cases were identified, with a date of symptom onset ranging between 23 February and 25 May 1951. The outbreak was detected on 24 April and by 29 April control measures, such as additional mass vaccination and isolation of cases and contacts, had been implemented. No individual details on vaccination status were available, but the vaccination coverage in Tilburg in 1951 was approximately 88% [14, 15]. The cases occurring before April 29 were traced back and their contacts identified. These efforts led to a very complete documentation of this outbreak. For almost all cases (n=50), it is known by whom or how they were most likely infected. Contacts of cases were also registered and monitored. The data are given in eTable 5.1; http://links.lww.com/EDE/B290. The exposure type was categorized as several types: family/household, friend/neighbor, work-related, and medical/religious setting. For some family/household cases or contacts, the date and duration of exposure were not known exactly. We therefore assumed that the duration was 21 days forward from the symptom-onset date of the related case. If the related case was put in isolation, we assumed contact exposure from the symptom-onset date of the related case to the isolation date. To demonstrate our tool in this paper, we analyzed the data as

they would have been observed on 30 May 1951. For the remainder of the paper we will call this date “the current date”. Ethical approval was deemed not necessary, as the data was already publicly available and does not allow identification of individual persons.

Monitoring period

An important decision in outbreak control concerns which contacts need follow-up and which can be set aside. For each contact *i* we define the monitoring period as the interval from the 5-percentile time-point to the 95-percentile time-point of the effective incubation period. The effective incubation period $f_{eff,i}(t)$ is defined as the total period from a first exposure to symptom onset. For any infected individual, the effective incubation period is the sum of the time from first exposure to effective transmission, and from effective transmission to symptom onset. The distribution of the effective incubation period is the convolution of the uniform distribution from first exposure to last exposure with the lognormal distribution that describes the incubation period distribution. In our method, repeated exposures are merged to one exposure interval, which is assumed to be uniformly distributed. For smallpox, the lognormal distribution that best fits the observed incubation periods has a median of 13.0 days and a dispersion of 1.13 days [16]. We denote the corresponding cumulative probability distribution function as $F_{eff,i}(t)$ and the corresponding survival distribution as $S_{eff,i}(t)= 1- F_{eff,i}(t)$. In the visual output, the monitoring period is displayed as a horizontal black line, and it is shown only for those contacts that are still monitored, i.e. with $S_{eff,i}$(current date) > 0.05.

Probability of infection for a contact

Each contact of exposure type *x* has a probability π_x of being infected and a probability of $(1- \pi_x)$ of not being infected. If not infected, the contact does not develop any symptoms. If infected, the contact will develop symptoms, and the time from first exposure to symptom onset will follow the effective incubation period. This defines a mixture distribution of a contact *i* of type *x* for its time to infection: $f_{x,i}(t) = \pi_x f_{eff,i}(t)$. We denote the corresponding cumulative probability distribution function as $F_{x,i}(t) = \pi_x F_{eff,i}(t)$ and the corresponding survival distribution as $S_{x,i}(t) = 1 - \pi_x F_{eff,i}(t)$. For each contact *i* who has a link of exposure type *x* to its linked case *j* and has not developed any symptoms up to time *t* since exposure, the probability that it is infected is calculated by Bayes’ rule as

$$\Pr(\text{infected}|\text{no symptoms at } t) = \frac{\Pr(\text{no symptoms at } t|\text{infected})\Pr(\text{infected})}{\Pr(\text{no symptoms at } t)} \quad (1)$$

which is, based on the definitions above, given by

$$p_{i,j,x} = \frac{\hat{\pi}_x S_{eff,i}(t)}{S_{x,i}(t)} \quad (2)$$

We used maximum likelihood estimation to estimate π_x. More details can be found in Appendix A.

Attack rate by exposure type

The exposure type-specific AR, π_x, is defined as the probability of a contact being infected, among contacts of exposure type *x*. In general, the exposure type-specific ARs will tell public health professionals the type of exposure through which contacts are at the highest risk of developing symptoms. In the context of contact tracing, the AR measures the true positive rate, or precision, of the contact definition. A higher value indicates a more efficient contact tracing. The AR is separately estimated for every exposure type *x* having a total number of cases and contacts greater than 5.

Reproduction number

By evaluating the course of the effective reproduction number *R(t)* during the outbreak, public health professionals can assess changes in the effective transmissibility of the disease [7, 11]. There are various ways of estimating *R(t)* over time [7, 11, 17], but most assume that the links between the cases, who infected whom, are not known. The *R(t)* then needs to be inferred indirectly, whereas when using contact tracing data, the *R(t)* can be estimated directly. For each case *j* we calculate the expected number of secondary cases. The secondary cases consist of confirmed secondary cases, contacts that are or might be infected but have not yet shown symptoms, and orphans. Orphans are infected individuals for whom the infector's identity is unknown, though by assumption the infector is one of the known cases. To calculate the first term on the righthand side, we count the number of confirmed secondary cases of case *j* ($D_j(t)$). Confirmed secondary cases are defined as contacts that were followed up as a result of exposure to a case, and eventually developed symptoms within their monitoring period. To calculate the second term, we sum over the contacts of case *j* and weigh each contact by his/her probability of being infected by case *j* ($p_{i,j,x}(t)$). To calculate the third term, we sum over the number of orphans and weigh each by his/her probability of being infected by case *j* [17].

$$R_j(t) = D_j + \Sigma_i \Sigma_x p_{i,j,x}(t) + \Sigma_k \frac{w(t_k - t_j)}{\Sigma_l w(t_k - t_l)} \qquad (3)$$

Here, the index *i* refers to all contacts, the index *k* to all 'orphan' cases, and the index *l* to all contacts that could have infected the *k*th orphan case. The function *w* gives the generation interval distribution, where the generation interval is defined as the duration between symptom onset of an infected individual and symptom onset of his/her infector. For smallpox, we take the generation interval to follow a gamma distribution with a mean and variance of 17.7 days [18].

For any given period of time, we can calculate the average value of the reproduction number for cases who developed symptoms within that period. Of interest here are the period from start of the outbreak to the time the first case was detected (including all primary cases with a date of onset of disease before 24 April 1951 and all secondary cases with an exposure

date before 24 April 1951) and the period from implementation of control measures to the current date (including all cases with an exposure date between 29 April 1951 and 30 May 1951). To visualize the time course of the average reproduction number, we calculate for each time t a running average over a time window $(t - \tau, t)$ of length τ. We choose τ equal to the generation interval for the specific infection. For smallpox it is 17.7 days [18]. We estimate 95% nonparametric bootstrap percentile confidence intervals [11] around the reproduction numbers, which reflect the uncertainty in whether contacts in follow-up will become cases and the uncertainty regarding possible links to orphan cases. The intervals are estimated by repeating the following steps a thousand times: first, draw from a binomial distribution with probability $p_{i,j,x}$ for every contact in follow up whether it is going to be a case or not and for every possible link to an orphan case if this a true link or not; second, sum the total number of secondary cases per case (this includes confirmed secondary cases, contacts in follow up and orphan cases); and third, compute the mean of the total number of secondary cases per case. By repeating this a thousand times, a thousand means are estimated, from which the 2.5% and 97.5% percentiles define the 95% confidence interval around the *R(t)* in a certain time interval.

Algorithm for layout of the time-oriented transmission tree

To provide an overview of the links between cases and contacts and the evolution of the outbreak over time, we adapted the Reingold-Tilford tree-drawing algorithm [19] to obtain a time-oriented transmission tree (see further details in Appendix B). In this node-link tree, cases and contacts are represented as nodes, and the exposures between cases and contacts are represented as links. The nodes of the cases are placed according to the first day of symptom onset, and the nodes of the contacts are placed according to the lower level of the monitoring period.

Model assumptions

Underreporting

The model assumes that all cases and contacts have been observed. However, this is most often not the case. This underreporting might therefore introduce bias to our model. As intensive daily contact tracing is in place during an emerging infectious disease outbreak, we assume that the number of missed secondary cases is minimal and subsequently that this underreporting has a limited effect on the estimation of the effective reproduction number. Underreporting does not play a role in the estimation of the attack rates, as we have defined the attack rate as the proportion of infected individuals among the total number of traced individuals, instead of susceptibles.

Reporting delay

The model does not take into account reporting delay, which may introduce bias in our estimates of the effective reproduction number and the attack rates. As intensive contact

tracing is in place, we assumed that reporting delays are minimal as all cases and contacts are under strict monitoring.

Introductions versus orphans

The model assumes there was only one introduction (index case) in the community, and that cases not reported as a contact by anyone else (i.e. orphans), have their infector among the observed cases. We considered it more likely that a case with an unknown source originates from another observed case than that a new introduction of the emerging infectious disease occurred in the community.

Simulation study

To allow assessing the accuracy of the estimated attack rates and effective reproduction number, and to assess the performance of our tool for a larger outbreak, we performed a simulation study. We used the outbreaker package in R to simulate an outbreak three times as large as the smallpox outbreak in Tilburg, with similar parameter settings. We compared the estimates of our model that use information on cases in follow-up to correct for right censoring with reference estimates obtained by including all present and future information on the infections and their contacts (no right censoring) and with naïve estimates obtained by ignoring any current information on the contacts (no correction for right censoring). A detailed description of the simulation study can be found in Appendix C.

Visualization

To provide an overview of the outbreak, we display the contact tracing data in a multi-panel plot (Figure 5.1) using the above-mentioned parameters and algorithms. It includes a large plot with three horizontal panels plus a fourth panel at lower right.

In the large plot, the bottom panel contains an epidemic curve of the outbreak so far. The epidemic curve is displayed as a histogram of the number of cases per three days from the date of symptom onset of the first case to the current date. We take as a rule of thumb: the width of the bars in the epidemic curve is usually one quarter of the length of the incubation period.

The middle panel, which is vertically aligned with the epidemic curve, contains a line graph of the course of $R(t)$ with its 95% CI during the outbreak. In addition, for every case the observed number of secondary cases $D_j(t)$ is displayed by date of symptom onset of case j as a scatterplot, in which the size of the dots is proportional to the frequency of the number of secondary cases.

The top and main panel, vertically aligned with those below, contains the time-oriented transmission tree. We have used the algorithm described above to place the nodes

and links. The nodes representing cases in the tree are displayed as squares (men) or as diamonds (women) at the date of symptom onset. The nodes representing contacts in follow-up are displayed as dark red bars reflecting their monitoring period. In this way, it can easily be seen which contacts are still at risk of developing symptoms, i.e., where the period overlaps the current date. Nodes are labeled with the case or contact ID number. The links between nodes are depicted as colored lines reflecting the various exposure types (family/household, friend/neighbor, medical/religious, work); the line types differ for links between cases (solid) and case–contact links (dashed).

A solid vertical bar intersects these three panels to indicate the current date. The time-period after the current date is colored grey to indicate that the outcomes are still unknown for that period. Finally, the plot is annotated with dashed vertical lines that indicate important dates, such as the day when the first case was discovered and the day when control measures were installed. With the addition of such annotations, it can easily be seen if the situation has changed after a certain time-point.

The fourth panel, lower right, contains a bar chart representing the exposure type-specific ARs and their 95% CIs. The bar colors match the line colors in the network plot. Exposure types are displayed only when their total number of cases and contacts is greater than 5. Finally, above the fourth panel is information containing some summary statistics of the outbreak so far (number of cases/contacts, number of contacts in follow-up or discarded, and maximum number of secondary cases per case) as well as important dates and explanation of the above-mentioned features.

The R 3.1.0 statistical software is used to construct this multi-panel plot; main packages used are *igraph*, *ggplot* and *shiny*.

Results

An overview of the outbreak situation at 30 May 1951 in a multi-panel plot is given in Figure 5.1. The runtime of the analysis of this outbreak at this date was 10.92 seconds. An overview of the outbreak at an earlier time, the 29th of April 1951, can be found in Appendix D.

Between 23 February and 30 May 1951, 52 smallpox cases including 28 men and 24 women were identified, with the highest number of cases reported at the beginning of May, as seen in the epidemic curve (bottom left panel, Figure 5.1). For three cases (ID 1, 34, 50), no source could be identified. By May 30, known as the current date, 22 contacts were in follow-up (i.e., within their surveillance period) and 174 contacts were discarded, as their surveillance was over. The latter are not shown to prevent crowding of the plot. If no new cases and contacts present, follow-up of all contacts was to end by mid-June 1951.

The time-oriented transmission tree (top left panel, Figure 5.1) shows that most cases were infected through household/family contact (31/52) (red lines), followed by friend/neighbor contact (9/52) (blue lines). The exposure type-specific ARs are depicted in the bar chart in the lower right panel of Figure 5.1. The AR was highest among those infected by work-related contact (AR= 42%; 95% CI: 9 - 81%) and by medical contact (AR= 31%; 95% CI: 5 – 71%).

The evolution of *R(t)* during the outbreak (left middle panel, Figure 5.1) shows several peaks in March and April, after which the *R(t)* declines to zero at the beginning of May. At the end of May, a slight increase in the *R(t)* is observed, indicating probable secondary cases among the contacts still in follow-up. The maximum number of secondary cases per case was 13 by case 35, who could be identified as the main spreader in this outbreak thus far. The transmission tree and the course of *R(t)* over time indicate that before the first case was discovered and additional control measures were implemented, the *R(t)* was decreasing. The *R(t)* before the first case was detected on 24 April was 2.42 (95% CI: 2.4, 2.53), and the *R(t)* after additional control measures were installed was 0.83 (95% CI: 0.43, 1.29). The *R(t)* before outbreak detection is slightly lower than the reproduction numbers reported in previous studies [20-22], however, when taking into account that the population in this outbreak was only partially susceptible due to childhood vaccination, the findings are quite comparable.

The simulation study in Appendix C shows that a larger outbreak can be studied in a similar way as the outbreak used in this study. The proposed estimators for the attack rate and effective reproduction number that include information on contact tracing to account for right censoring of the data, come closer to the actual values in comparison with another (naïve) estimator that does not correct for right censoring. The runtime of the analysis of the simulated outbreak was 27 seconds on a standard laptop computer (Intel Core i5 processor).

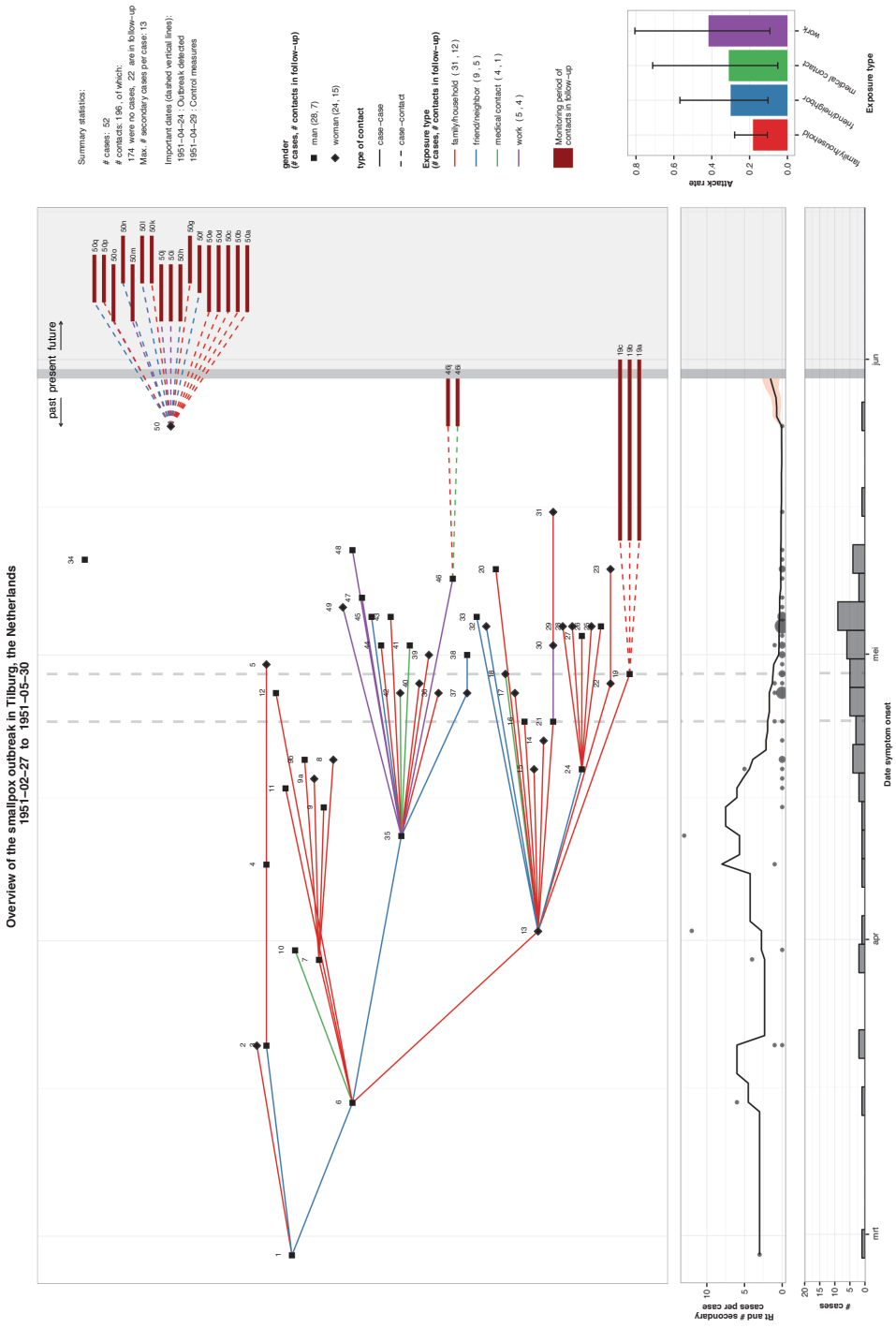
Overview of the smallpox outbreak in Tilburg, the Netherlands
1951–02–27 to 1951–05–30
past present future
Summary statistics:
cases: 52
contacts: 196, of which:
174 were no cases, 22 are in follow–up
Max. # secondary cases per case: 13
Important dates (dashed vertical lines):
1951–04–24 : Outbreak detected
1951–04–29 : Control measures
gender
(# cases, # contacts in follow–up)
man (28, 7)
woman (24, 15)
type of contact
case–case
case–contact
Exposure type
(# cases, # contacts in follow–up)
family/household (31 , 12)
friend/neighbor (9 , 5)
medical contact (4 , 1)
work (5 , 4)
Monitoring period of contacts in follow–up
Attack rate
Exposure type
family/household
friend/neighbor
medical contact
work
Rt and # secondary cases per case
cases
Date symptom onset
mrt
apr
mei
jun

Figure 5.1. Screenshot of the contact tracing visualization tool with an overview of contact tracing information during an outbreak of smallpox in Tilburg, the Netherlands, using information as of May 30, 1951. Top panel: Node-link diagram of the (possible) transmission events between cases and contacts over time. The cases are shown as symbols (square=men, diamond=women, circle= unknown) by date of onset of symptoms. The contacts with Seff,i(current date) > 0.05 are displayed as a horizontal dark red bar indicating their monitoring period. The links between cases (case-case links, solid) and between cases and contacts in follow-up (case–contact links, dashed) are depicted as lines and colored by exposure type. Middle panel: Course of R(t) during the outbreak and observed number of secondary cases per primary case. The course of the average case reproduction number, R(t) (with 95% nonparametric bootstrap percentile confidence intervals), was estimated using a left-sided moving average with a time-window of 18 days (the generation interval). The observed numbers of secondary cases per primary case are displayed as black dots, with the size of the dots being proportional to the frequency of the number of secondary cases. Bottom panel: The epidemic curve of the outbreak per three days by date of onset of symptoms. Bar chart: The exposure type-specific attack rates with their confidence intervals. Only exposure types with >5 cases/contacts are displayed.

Discussion

We have proposed a tool and methods to accurately infer outbreak parameters in real time from contact tracing data in order to enhance decision-making and, subsequently, the efficacy and efficiency of outbreak control. The tool provides an indication of the work-load of outbreak investigators (the number of contacts in follow-up) and gives an impression of the effective transmissibility of the disease during the outbreak (*R(t)*). Visualization of this information shows clearly which contacts are still at risk of developing symptoms and which are no longer at risk and can therefore be discarded, reducing the work-load. The estimation and visualization of exposure type-specific ARs helps public health practitioners to identify the exposure types through which contacts are at the highest risk of developing symptoms. Contact definitions can then be made more specific during the outbreak. In our example of smallpox, we surprisingly found that work-related and medical contacts yielded the highest AR, although family/household contact is usually more intense. Finally, the information available through contact tracing allows assessment of interventions. In the smallpox outbreak, it was observed that before the first case was discovered and the implementation of additional control measures, the *R(t)* was decreasing. This decrease might be attributed to the intervention, as cases identified just before its installation (up to 18 days, the generation interval) were less likely to cause secondary cases. The course of *R(t)* is likely to be consistent with effectiveness of the control measure.

Our approach has three main strengths. First and foremost is its ability to analyze the data in real time. Incubation period distributions are used to estimate when a contact is most likely to develop symptoms, providing an up-to-date overview of the contacts in follow-up and the related work-load for outbreak investigators. We added to the real-timeliness of the tool by taking the censored data into account. This involved estimating the probability of infection for a contact, using maximum likelihood estimation for those contacts still in follow-up to estimate the AR and *R(t)* during the course of the outbreak. The AR is often obtained by using aggregated numbers for total number of cases and contacts at a single time-point. Contacts in follow-up are usually neglected and left out of the analysis. However, the contacts in follow-up can still become cases, and the ARs might therefore be underestimated. Also, *R(t)* is often underestimated, as potential secondary cases are not taken into account. By accounting for the censoring of contacts still in follow-up, we could make our AR and *R(t)* estimations more accurate (as we have shown in our simulation study in Appendix C).

The second strength of our approach is that it provides a comprehensive overview of the outbreak at a glance. This graph is easily brought to table when outbreak management decisions are discussed, making it easier to use by outbreak investigators compared to other more software-oriented tools [8, 12, 13]. The tool is provided with a user-friendly interface, making the use of the tool relatively straightforward. Value has been shown for all

elements of the visual, such as the epidemic curve, the contact network, and the statistics; but having them readily available in one visual output with the latest results is novel. Our combination of elements offers a great advantage over previous methods focused on only one element of the outbreak analysis [7-9].

The third strength is that the output is presented visually. We specifically aimed for this, as the human mind can acquire more information through vision than through all of the other senses combined [23]. Besides fast and efficient processing, visualization has several other advantages [23], allowing the viewer to comprehend large amounts of data and perceive patterns. Problems with the data can become more easily apparent. In our smallpox example, it is immediately clear that the generation time between case 6 and 12 is improbably long and most likely due to a registration error or missed case in between. Finally, visualization facilitates hypothesis generation. For these reasons, we support the use of visuals instead of tables and text to communicate contact tracing information for decision- making. In addition, the visual output bridges the gap between the more software oriented-tools [8, 12, 13] and public health professionals whose programming skills are not always advanced.

Several issues regarding use of our tool in daily practice should be kept in mind. First, the visual is only as good as the input information. For this study, we used historical data of a well-documented outbreak. We are aware that during an outbreak, registration of all cases and contacts in a tidy way might not be the first priority. However, we expect that our approach to visualize the valuable information for decision-making provides an incentive for outbreak investigators to organize the data in the required format, as it will be time-saving in the end. On the other hand, since this method is so dependent on data quality, it can be used to trace errors of data entry as well. Second, the visual layout can be adapted to deal with very large outbreaks with numerous contacts, such as the pertussis outbreak in the Netherlands in 2012 [24], infectious diseases with a potentially long incubation time, such as tuberculosis [25], or other than human-to-human transmissible diseases. For very large outbreaks, the tool could be adapted by splitting the outbreak into smaller clusters. For other than human-to-human transmissible diseases, like environmental point-source outbreaks, the tool could be adapted by estimating the attack rates and transmission tree by environmental source. The estimates of the effective reproduction number are not relevant in such a situation.

There are various ways in which the tool can be improved upon in future studies. We chose to use the incubation period distribution and generation time distribution from the literature. One could also try to estimate these parameters from the data itself, with as a major advantage a higher accuracy of the estimates for that specific outbreak. However, the estimates would be very unstable at the beginning of the outbreak, as only limited

data is available. It could be a possible direction for future development of the tool to combine prior information from literature with a growing amount of data from an ongoing outbreak. Second, outbreaks are typically described by the time, pathogen type, place, and person characteristics. This tool was mainly focused on the dimensions of time and person. Future work on how to add the dimensions of pathogen type and place to this layout would be useful. A final minor point for improvement is the tree plotting algorithm, for which we adapted a proven algorithm to optimize the use of available space (Appendix B), but which in some cases results in crossing lines.

To conclude, we have proposed a tool to visualize contact tracing information to increase timely access to information on the characteristics of an emerging outbreak. The tool obtains and visualizes accurate outbreak parameters by taking into account the censoring of the data. The resulting information can help public health professionals to get a real-time overview of the outbreak and subsequently make better decisions in order to control the outbreak. Further research should focus on improving the network layout algorithm and enhancing the visual with data on pathogen type and place that are obtained during the outbreak.

References

1. Eames KT, Keeling MJ. Contact tracing and disease control. Proc Biol Sci. 2003;270(1533):2565-71.
2. Klinkenberg D, Fraser C, Heesterbeek H. The effectiveness of contact tracing in emerging epidemics. PLoS One. 2006;1:e12.
3. Kiss IZ, Green DM, Kao RR. Disease contact tracing in random and clustered networks. Proc Biol Sci. 2005;272(1570):1407-14.
4. Bonacic Marinovic A, Swaan C, van Steenbergen J, Kretzschmar M. Quantifying reporting timeliness to improve outbreak control. Emerg Infect Dis. 2015;21(2):209-16.
5. Fraser C. Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic. PLoS One. 2007;2(8):e758.
6. Jewell CP, Roberts GO. Enhancing Bayesian risk prediction for epidemics using contact tracing. Biostatistics (Oxford, England). 2012;13(4):567-79.
7. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178(9):1505-12.
8. Obadia T, Haneef R, Boelle PY. The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. BMC Med Inform Decis Mak. 2012;12:147.
9. Ghani AC, Donnelly CA, Cox DR, Griffin JT, Fraser C, Lam TH, et al. Methods for estimating the case fatality ratio for a novel, emerging infectious disease. Am J Epidemiol. 2005;162(5):479-86.
10. Cauchemez S, Boelle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, et al. Real-time estimates in early detection of SARS. Emerg Infect Dis. 2006;12(1):110-3.
11. Hens N, Calatayud L, Kurkela S, Tamme T, Wallinga J. Robust reconstruction and analysis of outbreak data: influenza A(H1N1)v transmission in a school-based population. Am J Epidemiol. 2012;176(3):196-203.
12. Jombart T, Aanensen DM, Baguelin M, Birrell P, Cauchemez S, Camacho A, et al. OutbreakTools: a new platform for disease outbreak analysis using the R software. Epidemics. 2014;7:28-34.
13. Höhle M. An R package for the monitoring of infectious diseases. Computational Statistics. 2007;22(4):571-82.
14. Sas GJ. Klinische en epidemiologische waarnemingen tijdens de pokkenepidemie te Tilburg in 1951: Universiteit Leiden; 1954.
15. Ministerie van sociale zaken en volksgezondheid. Rapport omtrent de pokkenepidemie te Tilburg in 1951. Staatsdrukkerij- en uitgeverijbedrijf, 's-Gravenhage: 1953 33565.
16. Sartwell PE. The incubation period and the dynamics of infectious disease. Am J Epidemiol. 1966;83(2):204-6.
17. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 2004;160(6):509-16.
18. Vink MA, Bootsma MC, Wallinga J. Serial intervals of respiratory infectious diseases: a systematic review and analysis. Am J Epidemiol. 2014;180(9):865-75.
19. Reingold EM, Tilford JS. TIDIER DRAWINGS OF TREES. IEEE Transactions on Software Engineering. 1981;SE-7(2):223-8.
20. Gani R, Leach S. Transmission potential of smallpox in contemporary populations. Nature. 2001;414(6865):748-51.
21. Eichner M, Dietz K. Transmission potential of smallpox: estimates based on detailed data from an outbreak. American journal of epidemiology. 2003;158(2):110-7.
22. Riley S, Ferguson NM. Smallpox transmission and control: spatial dynamics in Great Britain. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(33):12637-42.
23. Ware C. Information Visualization: Perception for Design: Elsevier Science Morgan Kaufmann; 2004.
24. van der Maas NA, Mooi FR, de Greeff SC, Berbers GA, Spaendonck MA, de Melker HE. Pertussis in the Netherlands, is the current vaccination strategy sufficient to reduce disease burden in young infants? Vaccine. 2013;31(41):4541-7.
25. Borgdorff MW, Sebek M, Geskus RB, Kremer K, Kalisvaart N, van Soolingen D. The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. Int J Epidemiol. 2011;40(4):964-70.

Appendix A

MAXIMUM LIKELIHOOD ESTIMATION FOR PROBABILITY ON INFECTION

Notation and data

By a contact we mean an individual that has been identified in contact tracing, irrespective of the fact whether this individual is infected or not. We use the subscript $i = 1, 2, \ldots, n$ to label the contacts. Contacts are categorized by their type (household, work etc). We use the subscript x to denote type.

By an orphan we mean an infected individual for which the infector hasn't been traced.
For each contact i we have observation on time of start of exposure, $t_{start,i}$, and the end of exposure $t_{end,i}$. We have a censoring indicator d_i which is 0 if the contact has not shown any symptoms at the current time t, and which is 1 if the contact has shown symptoms before the current time t. For a contact i who has shown symptoms before the current time t ($d_i = 1$), we denote the time of symptom onset as $t_{sym,i}$. For a contact i who has not shown symptoms the current time, we denote the current time as censoring time for $t_{sym,i}$. For each contact i we calculate the time from start of exposure to the time of symptom onset / censoring time:

$$t_i = t_{sym,i} - t_{start,i} \tag{1}$$

Exposure and incubation period

The incubation period is defined as the time from exposure to symptom onset. For many infectious diseases this distribution is well described by a lognormal distribution. For smallpox the best fitting lognormal distribution to observed incubation periods has a median of 13 days and a dispersion of 1.13 (Sartwell, 1966). We denote this distribution as f_{inc}:

$$f_{inc}(t) \sim lognormal(ln(median),\ ln(dispersion)^2) \tag{2}$$

We denote the cumulative distribution as F_{inc} and the corresponding survival distribution as $S_{inc} = 1 - F_{inc}$. For most cases we do not have a single moment of exposure, but a period from first exposure to last exposure. We assume that effective transmission from the infector to the infected can occur with equal probability during the period from first to last exposure. We denote this exposure distribution as f_{exp}.

$$f_{exp,i}(t) \sim uniform(t_{start,i},\ t_{end,i}) \tag{3}$$

We define an effective incubation period as the total period from a first exposure to symptom onset. For any infected individual, the effective incubation period is the sum of the time from first exposure to effective transmission, and from effective transmission to symptom onset. The distribution of the effective incubation period is the convolution of the uniform

distribution from first exposure to last exposure with the lognormal distribution that describes the incubation period distribution. We denote the effective incubation period distribution as

$$f_{eff,i}(t) = f_{inc}(t) * f_{exp,i}(t) \quad (4)$$

We denote the corresponding cumulative probability distribution function by $F_{eff,i}(t)$, and the corresponding survival distribution as $S_{eff,i}(t) = 1 - F_{eff,i}(t)$.

Probability of infection for a contact

Each contact of type x has a probability π_x of being infected and a probability $(1- \pi_x)$ of not being infected. If not infected, the contact does not develop any symptoms. If infected, the contact will develop symptoms and the time from first exposure to symptom onset will follow the effective incubation period. This defines a mixture distribution of a contact of type x for its time to infection:

$$f_{x,i}(t) = \pi_x f_{eff,i}(t) \quad (5)$$

$$F_{x,i}(t) = \pi_x F_{eff,i}(t) \quad (6)$$

with a corresponding survival distribution

$$S_{x,i}(t) = 1 - \pi_x F_{eff,i}(t) \quad (7)$$

We can calculate the type-specific hazard of symptom onset as

$$h_{x,i}(t) = \frac{\pi_x f_{eff,i}(t)}{1-\pi_x F_{eff,i}(t)} \quad (8)$$

Likelihood equation for probability of infection

The likelihood contribution of the ith contact with symptom onset/censoring time t_i and indicator d_i is

$$L_i(\pi_x|t_i, d_i) = [h_{x,i}(t_i)]^{d_i} S_{x,i}(t_i) \quad (9)$$

The corresponding contribution to the log likelihood is

$$\ell_i(\pi_x|t_i, d_i) = d_i \ln h_{x,i}(t_i) + \ln S_{x,i}(t_i) \quad (10)$$

The total log likelihood is

$$\ell(\pi_x|t,d) = \Sigma_i \ell_i(\pi_{x,i}|t_i, d_i) = \Sigma_i d_i \ln h_{x,i}(t_i) + \Sigma_i \ln S_{x,i}(t_i) \quad (11)$$

substituting the definitions for $h_{x,i}$ and $S_{x,i}$ and simplifying gives

$$\ell(\pi_x|t,d) = c + \Sigma_{i:d_i=1} \ln \pi_x + \Sigma_{i:d_i=0} \ln (1 - \pi_x F_{eff,i}(t_i)) \quad (12)$$

where c is a constant. For practical purposes we might simplify this equation further. We have n contacts, of which a number of n_s contacts showed symptoms, a number of n_f contacts are in follow up as they might be incubating infection, and a number of n_a contacts that showed no symptoms during their monitoring period.

$$\ell(\pi_x|t,d) = c + \mathrm{n_s} \ln (\pi_x) + n_a \ln(1 - \pi_x) + \Sigma_{i=1}^{i=n_f} \ln (1 - \pi_x F_{eff,i}(t_i)) \quad (13)$$

We can check that when the epidemic is over, and there are no contacts incubating infection, this reduces to the binomial likelihood for proportion symptomatic.

Estimator for probability of infection

The maximum likelihood estimate for the probability of contact of type x, π_x, to be infected can be found by maximizing the log-likelihood.

$$\hat{\pi}_x = argmax_{\pi x} \ell(p_{x,i}|t,d)) \quad (14)$$

The 95% confidence interval for probability of contact of type x to be infected, π_x, is obtained as a profile likelihood. We include values of π_x where the deviance $2(\ell(\hat{\pi}_x) - \ell(\pi_x)$ is smaller than 3.841, the critical value for Pearson's chi-square test at a level of significance of 0.05 on one degree of freedom.

Probability of infection for a contact of type *x* that has not developed any symptoms up to time *t*

For each contact i which has a link of type x to its infector case j, the probability that it is infected at time t is given by

$$p_{i,j,x} = \frac{\hat{\pi}_x S_{eff,i}(t)}{S_{x,i}(t)} \quad (15)$$

Numerical considerations

We approximate the cumulative effective incubation period distribution $F_{eff,i}(t)$ by drawing a sample from the uniform distribution, drawing a sample from the lognormal distribution, and adding these to obtain a sample from the effective incubation period distribution. We repeat this 10,000 times, and rank the samples. We obtain a close approximation to the probability

density of the effective incubation period distribution, $F_{eff,I}(t)$. We evaluate the log-likelihood for values of π_x ranging from 0 to 1. For these values we identify first the maximum likelihood value, and then the values that lie within the 95% confidence interval around the maximum likelihood value.

Appendix B

THE LAYOUT-ALGORITHM OF THE TRANSMISSION TREE

We adapt the Reingold-Tilford tree drawing algorithm to get a time-oriented transmission tree. In this appendix, we explain the adaptations step by step, and we illustrate the adaptations with a hypothetical tree structure.

We start with placing the cases and contacts in a standard node-link diagram (Fig. B1A). We apply the Reingold-Tilford tree drawing algorithm on this diagram (Fig. B1B). This algorithm is based on a few principles: 1) nodes at the same generation of the tree should lie along a straight line (blue dotted lines), and the straight lines defining the generations should be parallel, 2) a left son should be positioned to the left of its father and a right son to the right, and a father should be centered over its sons. This algorithm results in a generation-based transmission tree. The next step is rotating the positions of the nodes such that cases in the same generation are aligned at the same horizontal position (Fig. B1C). We could shift the horizontal position of a case towards the left or the right such that this horizontal position reflects time of symptom onset and accordingly we could shift the horizontal position of a contact such that it becomes the time window for the 90%-interval of sympom onset; unfortunately, this would result in several crossing links (Fig. B1D). We introduce a trick to reduce the number of crossing links: before we reposition the nodes according to calendar time, we add ‘dummy offspring’ in such a way that all branches of the transmission tree have at least one node left in the last generation of the transmission tree (Fig. B1E). In the example, we created three generations of ‘dummy offspring’ for nodes E, F and H, which are in the same and last generation as nodes I, J, K, L and M of the transmission tree. We now reposition the nodes according to calendar time (Fig. B1F), and remove the ‘dummy offspring’ (Fig. B1G).

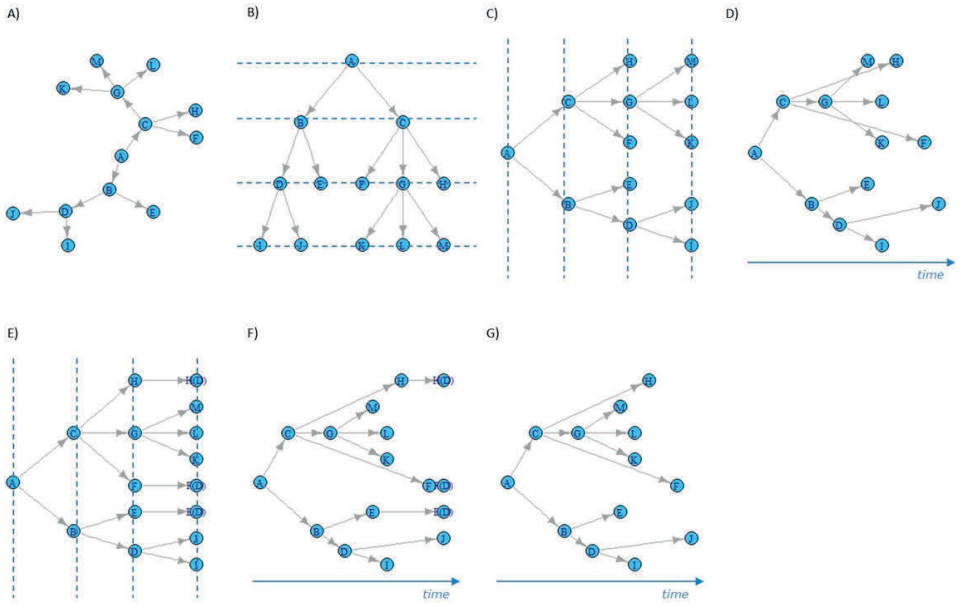

Figure B1: Steps in creating the layout of the transmission tree. A) Place the cases and contacts in a standard node-link diagram. B) Apply the Reingold-Tilford algorithm on the data. C) Align the generations vertically. D) Reposition the horizontal position of the nodes according to calendar time. E) To reduce the number of crossing links, add 'dummy offspring' to the vertically aligned transmission tree. F) Reposition the nodes and 'dummy offspring' according to time. G) Remove the'dummy offspring'.

Appendix C
SIMULATION STUDY

In order to assess the accuracy of the estimator of the attack rates (AR) and the effective reproduction numbers we performed a simulation study. We used the R package 'Outbreaker' and the simOutbreak function to generate an outbreak of about three times the size of the outbreak used in the main paper. We generated an outbreak in a population of 150 individuals for an infection with a basic reproduction number of 3, starting with one infectious case on March 1, 2015. We simulated contacts by linking each simulated case to a random number of contacts, this number was drawn from a discrete uniform distribution *Unif*{3, 10}. We simulated the exposure dates, durations of exposure and exposure routes for each contact. The code for generating the data is given at the end of this appendix.

An overview of the simulated outbreak for May 30, 2015 is shown in Figure C1. The epidemic curve, the time course of the reproduction number and the histogram with attack rates present a summary of the simulated data that are available at that point in time. The network gives a clear picture of how the infection was spreading at the start of the outbreak, although the more than 300 contacts in follow-up appear crowded.

The proposed estimator of the overall attack rate uses information on cases in follow-up to correct for right censoring. We compare the performance of this proposed estimator to two alternative estimators. One alternative estimator, the reference estimator, is obtained by including all present and future information on the infections and their contacts (no right censoring). Another estimator, the naive estimator, is obtained by ignoring any current information on the contacts (no correction for right censoring). Figure C2 shows the comparison between the proposed estimator of the overall attack rate and the two alternatives. One can see that the estimate of the overall attack rate is very uncertain when only a few cases have been observed. As more cases are observed, the confidence interval shrinks and closes in on the actual value without right censoring. This actual value is covered by the confidence interval throughout the epidemic. The proposed estimator is consistently better than the naive estimator that ignores information on contacts in follow-up and that does not correct for right censoring of the data.

The proposed estimator of the effective reproduction number also uses information on cases in follow- up to correct for right censoring. We again compare the performance of this proposed estimator to two alternative estimators. Again a reference estimator, that is obtained by including all present and future information on the infections and their contacts (no right censoring), and a naive estimator, that is obtained by ignoring any current information on the contacts (no correction for right censoring). Figure C3 shows the comparison between the proposed estimator of the effective reproduction number and the

two alternatives at four points in time during the outbreak. One can see that the estimate of the effective reproduction number is very uncertain when only a few cases have been observed at the beginning of the outbreak. As more cases are observed, the confidence interval shrinks. The proposed estimator is performing considerably better than the naive estimator that consistently underestimates the effective reproduction number in the most recent time intervals.

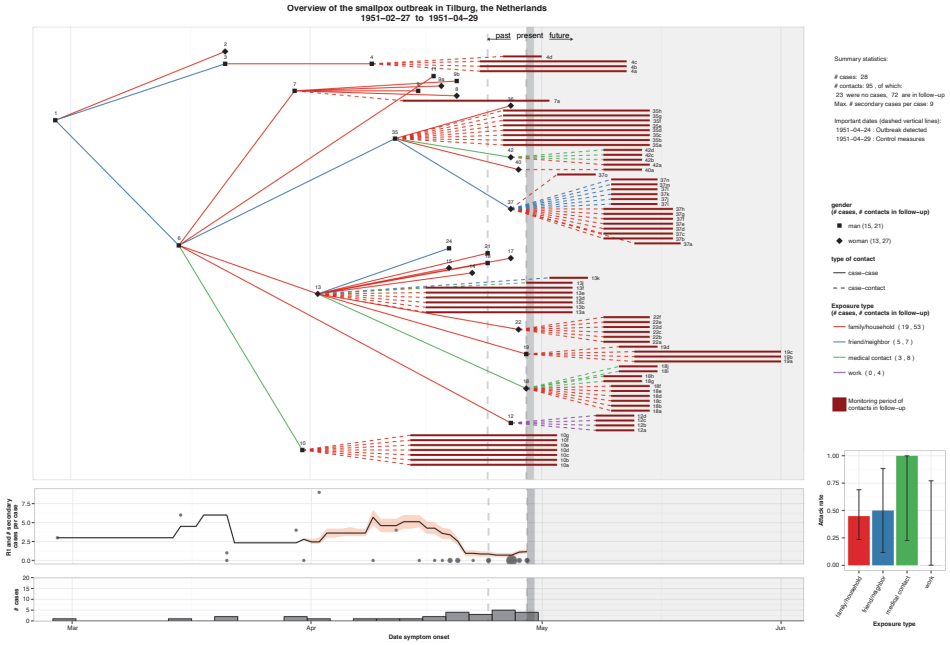

Figure C1. Overview of the simulated outbreak at May 30, 2015

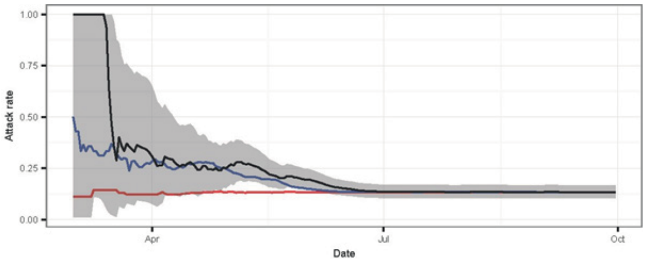

Figure C2. Estimators of the overall attack rates over time. Black: estimator of the overall attack rate over time that uses information on cases in follow-up to correct for right censoring (with 95% CI in grey); red: naive estimator of the overall attack rate over time that is obtained by ignoring any current information on the contacts (no correction for right censoring); blue: reference estimator of the overall attack rate over time that is obtained by including all present and future information on the infections and their contacts (no right censoring).

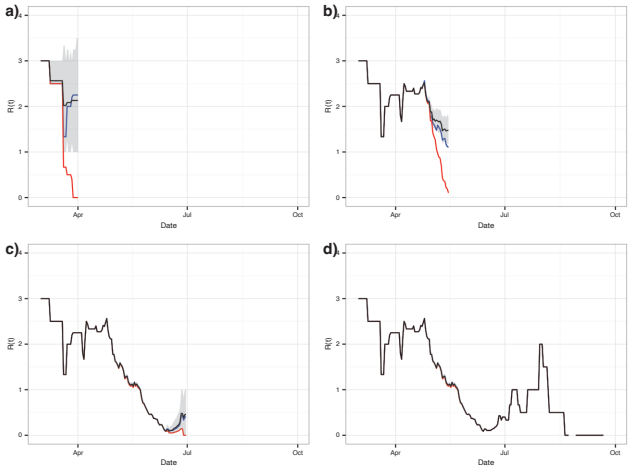

Figure C3: Estimators of effective reproduction numbers over time. a) Situation at 1 April 2015, b) Situation at 15 May 2015, c) Situation at 30 June 2015, d) Situation at 30 September 2015. Black: estimator of the effective reproduction number that uses information on cases in follow-up to correct for right censoring (with 95% CI in grey); red: naive estimator of the effective reproduction number that is obtained by ignoring any current information on the contacts (no correction for right censoring); blue: reference estimator of the effective reproduction number that is obtained by including all present and future information on the infections and their contacts (no right censoring).

```
R code to generate simulated dataset :
# Simulation outbreak
library(outbreaker)
# Simulate data using the outbreaker package
set.seed(10)
dat <- list(n=0)
## simulate data with at least 20 cases
while(dat$n < 20 ) {
dat <- simOutbreak(R0 = 3,  infec.curve = c(dgamma(seq(from =1 ,  to= 40 ,  by=1) ,
17.7)) , n.hosts=  150, duration = 250)}
simdat<- data.frame(ID= dat$id, IDSource=  dat$ances ,  date= dat$onset)
# add additional data for cases
simdat$Type<- c ("Case")
route <- c (rep ("Family/Household" , 31 ), rep("Friend/neighbor", 9),
rep("Medical", 4), rep("Work", 5)) simdat$Exposuretype <- sample(route,
length(simdat$Type), replace= T)
expdurfam <- rgamma (1000, 4)
simdat$Exposureduration <- c ( )
for(i in 1: length(simdat$Exposuretype)){
if(simdat$Exposuretype[i]== "Family/Household") {
simdat$Exposureduration[i]<- round(sample(expdurfam, 1), 0) } else {
simdat$Exposureduration[i]<- 1}}
incperdist<- rlnorm(1000, log(13), log(1.13))
simdat$Expdate<- c( )
for (i in 1: length(simdat$date)){
simdat$Expdate[i]<- simdat$date[i] - round(sample(incperdist, 1), 0)-
simdat$Exposureduration[i]}
#add random number (between 3 and 10) of contacts to every case
for (i in 1 : length(simdat$ID)){
ncontacts <- sample(c(3:10), 1)
id <- c(paste(c(rep(i, ncontacts)), letters[1: ncontacts], sep=""))
source <- c(rep(i, ncontacts))
datesample <- sample(c(-3:5), ncontacts, replace=T)
expdate <- simdat$date[i] + datesample
route <- c(rep("Family/Household", 12), rep("Friend/neighbor", 5), rep("Medical",
1), rep("Work", 4))
exposuretype <- sample(route, ncontacts, replace=T)
exposureduration  <- c()
for (j in 1: length(exposuretype)){
if (exposuretype[j] == "Family/Household"){
exposureduration[j] <- round(sample(expdurfam, 1), 0) } else {
exposureduration[j]<-1}}
contact.df <- data.frame(ID= id, IDSource= source, date= rep(NA, ncontacts),
Type= c(rep("Contact", ncontacts)), Exposuretype = exposuretype, Exposureduration=
exposureduration, Expdate= expdate)
simdat <- rbind(simdat, contact.df)
}
simdat$Gender <- sample(c("man", "woman"), length(simdat$Type), replace=T)
simdat$DOD <- as.Date("2015/03/01", origin= "1970-01-01") + simdat$date
simdat$Exposuredate <- as.Date("2015/03/01", origin= "1970-01-01") + simdat$Expdate
```

Appendix D

OVERVIEW OUTBREAK AT DATE OF IMPLEMENTATION OF CONTROL MEASURES

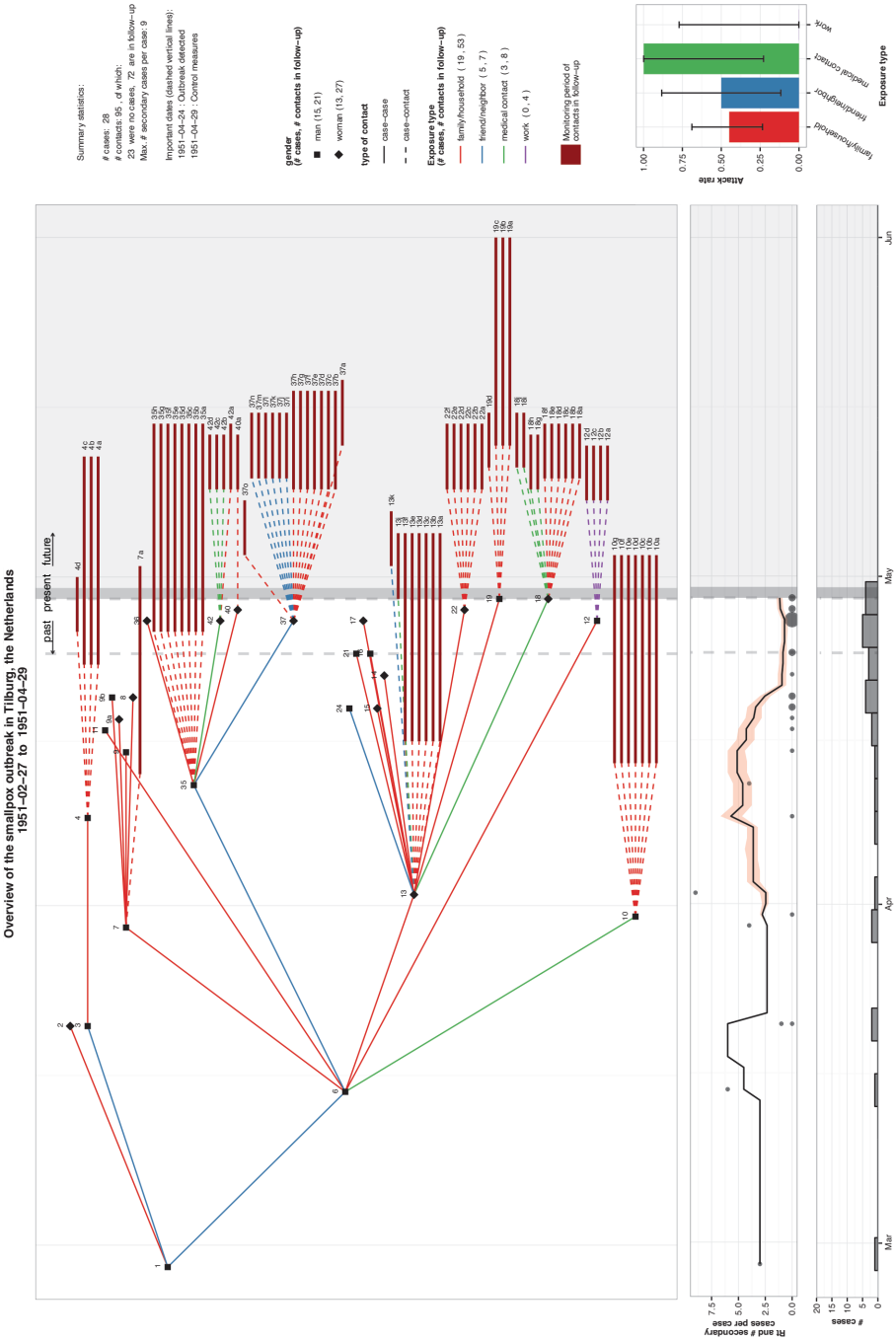

Figure D1: Overview of the smallpox outbreak in Tilburg, the Netherlands at April 29, 1951