Epidemics 17 (2016) 10–18

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Estimating infectious disease transmission distances using the overall distribution of cases

Henrik Salje[a,b,c,d,*], Derek A.T. Cummings[a,b,e,f], Justin Lessler[a,e]

[a] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA
[b] Mathematical Modeling of Infectious Diseases Unit, Institut Pasteur, Paris, France
[c] CNRS, URA3012, Paris 75015, France
[d] Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris 75015, France
[e] Department of Biology, University of Florida, Gainesville, FL, USA
[f] Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA

ARTICLE INFO

Article history:
Received 11 June 2016
Received in revised form 6 October 2016
Accepted 6 October 2016
Available online 7 October 2016

Keywords:
Transmission distance
Outbreaks
Infectious disease
Epidemiology
Foot-and-mouth disease

ABSTRACT

The average spatial distance between transmission-linked cases is a fundamental property of infectious disease dispersal. However, the distance between a case and their infector is rarely measurable. Contact-tracing investigations are resource intensive or even impossible, particularly when only a subset of cases are detected. Here, we developed an approach that uses onset dates, the generation time distribution and location information to estimate the mean transmission distance. We tested our method using outbreak simulations. We then applied it to the 2001 foot-and-mouth outbreak in Cumbria, UK, and compared our results to contact-tracing activities. In simulations with a true mean distance of 106 m, the average mean distance estimated was 109 m when cases were fully observed (95% range of 71–142). Estimates remained consistent with the true mean distance when only five percent of cases were observed, (average estimate of 128 m, 95% range 87–165). Estimates were robust to spatial heterogeneity in the underlying population. We estimated that both the mean and the standard deviation of the transmission distance during the 2001 foot-and-mouth outbreak was 8.9 km (95% CI: 8.4 km–9.7 km). Contact-tracing activities found similar values of 6.3 km (5.2km–7.4 km) and 11.2 km (9.5 km–12.8 km), respectively. We were also able to capture the drop in mean transmission distance over the course of the outbreak. Our approach is applicable across diseases, robust to under-reporting and can inform interventions and surveillance.

1. Introduction

Characterizing the spatial patterns of disease transmission is crucial to our understanding of pathogen dispersal. Public health interventions implicitly target next generations of transmission through contact tracing and spatial targeting of quarantine, isolation or other control measures, though often with crude information about where pathogens will move in space. More information about where cases may arise in relation to identified cases could help target resources both for control and enhanced surveillance. Despite its usefulness, the geographical mean distances between the locations of cases in relation to the individuals that infected them, have been difficult to elucidate. We rarely observe infection pairs (i.e., who infected whom) in a transmission network. Where only a minority of cases are observed, analyses tend to be restricted to characterizing the spatial and temporal scales at which cases tend to occur together but the relationship between spatial clustering and transmission distance is complex (Bhoomiboonchoo et al., 2014; Grabowski et al., 2014; Lin et al., 2011; Morrison et al., 1998; Salje et al., 2015, 2012). Only where we have been able to observe the majority of cases in a transmission network or we have detailed epidemiological data on who infected whom, has estimation of mean transmission distances previously been possible (Assiri et al., 2013; Ferguson et al., 2001a; Keeling et al., 2004).

It is not surprising that we are rarely able to reconstruct transmission pathways for outbreaks. Directly estimating the distance between sequential cases requires both the identification of cases and their infectors. Such contact tracing efforts can be expensive and time-consuming. In some cases it may be impossible. Usually only a fraction of cases are detected. Not everyone infected will develop symptoms severe enough to be detected (e.g., most dengue cases are not severe enough to seek care), and even the best

* Corresponding author at: Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.
E-mail address: hsalje@jhu.edu (H. Salje).

http://dx.doi.org/10.1016/j.epidem.2016.10.001

surveillance systems rarely identify 100% of symptomatic cases. Further, if there exists an intermediary vector or reservoir (such as the case of dengue, chikungunya or cholera), sequential cases in a transmission chain may never have been in contact with each other. Phylogeographic methods have been developed to estimate rates of viral movement across countries or continents under these conditions (Faye et al., 2015; Rabaa et al., 2013). However, these approaches have not yet been able to reliably capture micro-scale dynamics except in isolated settings such as hospital-based outbreaks (Cotten et al., 2013; Iles et al., 2014; Pybus et al., 2012; Rabaa et al., 2010), and may be impossible where genome mutation rates are particularly low or high relative to the generation time. Even where phylogenetic approaches can be used, it is likely to require potentially prohibitive labor-intensive sequencing of large numbers of pathogens throughout the course of an outbreak (Stack et al., 2010). Other fields have attempted to infer movement properties in poorly observed settings. Plant biology, for example, has developed methods to describe seed dispersal in situations where the source is unknown and thereby understand the relative importance of wind and animal movements in seed spread (Nathan and Muller-Landau, 2000). However, these methods have not been successfully applied to human disease spread.

Here, we present an approach to estimate the mean transmission distance in infectious disease processes using only the point locations of cases (e.g., place of residence), times at which individuals become symptomatic and the generation time distribution of the pathogen. The method is applicable in situations with full data as well as those where only a small proportion of infections are observed. We demonstrate the robustness of our approach using simulated data and then apply it to data from an outbreak of foot-and-mouth disease in the UK in 2001.

Box 1: Overview of key terms

Transmission linkage (θ)—The number of transmission events that link two cases (see example in Fig. 1)

Transmission kernel—The probability distribution function of the distance between sequential cases in a transmission chain

Most recent common ancestor (MRCA)—The most recent infector that can link a pair of cases

Mean transmission distance (μ_k)—The mean of the transmission kernel

Standard deviation of transmission distance (σ_k)—The standard deviation of the transmission kernel

Mean distance between θ transmission-linked pairs ($\mu_a(\theta,\mu_k,\sigma_k)$)—The mean distance between cases separated by θ transmission events where the transmission kernel has mean μ_k and standard deviation σ_k

Transmission-linkage weights ($w(\theta, t_1, t_2)$)—The proportion of case pairs where one occurs at t_1 and the other at t_2 that are separated by θ transmission events

Mean distance between all pairs ($\mu_t(t_1, t_2, \mu_k, \sigma_k)$)—The mean distance separating all pairs of cases where one occurs at t_1 and the other at t_2 and the transmission kernel has mean μ_k and standard deviation σ_k

Observed mean distance between case-pairs ($\mu_t^{obs}(t_1, t_2)$)—The observed mean distance separating all pairs of cases where one occurs at t_1 and the other at t_2

2. Methods

2.1. Distribution of distances between cases

In outbreaks originating from a single introduction into a community, a pair of cases occurring at time points t_1 and t_2 can be separated by a variable number of transmission events (denoted by θ, the number of infection events required to link a pair of cases) (Box 1 and Fig. 1). For example, two cases occurring at the same time may have been infected by the same infectious individual (in which case $\theta = 2$) or alternatively, their most recent common ancestor (MRCA) may be two or more generations back ($\theta > 2$). The distance between sequential cases in a transmission chain (i.e. $\theta = 1$) can be characterized by a transmission kernel, which we define here as the probability density function of all transmission distances during an epidemic. If we assume a constant isotropic transmission kernel (i.e. one with no directional preference), that transmission events are independent of each other and each infected individual has a single infector (i.e., co-infections do not occur), the distance between pairs of cases will depend on the number of transmission events that separate them. However, without detailed genetic information on the infecting pathogen or contact tracing information, we are unlikely to be able to directly identify the number of transmission events that separate any two cases. We can, however, calculate the mean distance between all observed pairs of cases that occur at two time points ($\mu_t^{obs}(t_1, t_2)$, the mean of the distribution represented by the solid black line in Fig. 1).

If we know the proportion of case-pairs at two time points that are separated by each possible θ, we can estimate the mean distance between all case pairs as a weighted sum:

$$\mu_t(t_1, t_2, \mu_k, \sigma_k) = \sum_i w\left(\theta = i, t_1, t_2\right) \cdot \mu_a\left(\theta = i, \mu_k, \sigma_k\right) \qquad (1)$$

where $\mu_t(t_1, t_2, \mu_k, \sigma_k)$ is the mean distance separating all pairs of cases where one occurs at t_1 and the other at t_2; $\mu_a\left(\theta, \mu_k, \sigma_k\right)$ is the mean distance between pairs of cases separated by θ transmission events where the transmission kernel has mean μ_k and standard deviation σ_k; and $w\left(\theta, t_1, t_2\right)$ are the weights representing the proportion of case pairs occurring at t_1 and t_2, respectively that are separated by θ transmission events. The variance of the distance between all case pairs can be similarly estimated (see Text S1).

We do not need to assume that the number of transmission events that separate a pair of cases infected at the same time is even (as would be the case if the generation time was of a fixed duration) or that individuals infected at the same time are from the same generation. Instead we can use information on the generation time distribution to calculate $w\left(\theta, t_1, t_2\right)$.

2.2. Estimation of weights

To estimate $w\left(\theta, t_1, t_2\right)$, we extended a method developed by Wallinga and Teunis that calculates the probability that a case occurring at time t_1 was infected by a case at time t_2 based on a known generation time distribution, *g(x)* and the number of cases occurring at each time point (Wallinga and Teunis, 2004). We produce an *n* x *n* matrix, where cell *[i, j]* represents the probability that a case *i* was infected by a case with the same time of disease onset as case *j* (the Wallinga-Teunis matrix) and *n* is the total number of cases. For each pair of cases, we can use the Wallinga-Teunis matrix to estimate the probability that they are separated by θ transmission events by multiplying together the cells of each unique chain (see Fig. 2 for a worked example). This assumes that the generation times for all infections were independent of each other and that only the day of symptom onset affected the probability of case *i* infecting case *j*. We could compute the probability of every possible path linking two cells, however, this quickly becomes computationally intractable. Instead we sampled transmission trees by randomly choosing the infector for each case. To do this we take each case in turn and randomly drew its infector out of all the other cases, with the probability of any other case being the infector coming from the Wallinga-Teunis matrix (i.e. determined by the time between the cases and the generation time distribution). Note that