Type and quantity of data needed for an early estimate of transmissibility when an infectious disease emerges

N G Becker (Niels.Becker@anu.edu.au)[1], D Wang[1], M Clements[1]
1. National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia

Citation style for this article:
Becker NG, Wang D, Clements M. Type and quantity of data needed for an early estimate of transmissibility when an infectious disease emerges. Euro Surveill. 2010;15(26):pii=19603. Available online: http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19603

Article published on 1 July 2010

An early estimate of disease transmissibility is essential for a well-informed public health response to a newly emerged infectious disease. In this study, we ask what type and quantity of data are needed for useful estimation of the initial reproduction number (*R*). It is possible to estimate *R* from case incidence data alone when the growing incidence of cases displays a wave pattern, because the pattern provides information about the serial interval (the time elapsed between the onset of symptoms of a case and symptom onset in individuals infected by that case). When the mode of the serial interval distribution is small, 1.5 days or less, there is generally no informative wave pattern in the observed series of daily incidences. The precision of the estimate of *R* is then improved substantially by having some observations on the serial interval. For an infectious disease with characteristics such as those of influenza, an estimate of *R* able to inform plans to mitigate transmission is obtained when the cumulative incidence of cases reaches about 300 and about 10 observations on the serial interval are available.

Introduction

Concern about the risk posed to humans by avian influenza A(H5N1) encouraged substantial planning for the possible emergence of pandemic influenza [1-4]. The emergence of the pandemic influenza A(H1N1) strain in 2009 further highlighted the importance of pandemic preparedness. Key elements of preparedness plans are disease transmissibility, the rate of disease progression and how these change with use of antiviral drugs, vaccines and social-distancing measures. Assumptions about disease transmissibility and progression are necessarily based on data from past pandemics and seasonally circulating influenza strains, but a future pandemic strain may have quite different characteristics.

Preparedness planning deals with this uncertainty by assessing the effectiveness of interventions under different scenarios. When a new viral strain emerges, it is important to determine which scenario obtains, because the effectiveness of some interventions is scenario dependent. For example, targeted use of antiviral drugs, early and liberally, may contain an influenza strain with a modest transmission rate, but would be ineffective against a highly transmissible strain. Timeliness is also important, because an intervention is most effective when introduced early. Here we consider what data from the early stage of an outbreak, and how much, are needed to inform decisions about interventions needed to mitigate the impact of a pandemic to a manageable level.

It is convenient to quantify disease transmissibility by *R*, the effective reproduction number of infective individuals. At any time, *R* is the mean number of infections generated by a 'typical' infective person, given current levels of immunity and public health interventions. It quantifies the growth in the number of cases from one generation of cases to the next. We aim to estimate the initial *R* from early incidence data of an outbreak. Incidence data alone seem inadequate for this estimation: we also need information about the serial interval, the time elapsed between the onset of symptoms of a case and symptom onset in individuals infected by that case. The artificial incidence series A and B of Table 1 illustrate this point. Comparing incidences on days 0, 2, 4, 6 and 8 suggests the two outbreaks are growing similarly over time, while comparing cumulative incidences suggests series B is the larger threat. However, series A actually poses the greater threat (larger eventual attack rate) because it is consistent with $R = 4$ and every infected person having a short symptomatic infectious period on the second day after infection, while series B is consistent with $R = 2$ and a short symptomatic infectious period on the first day after infection. In other words, reproduction numbers corresponding to incidences that appear to be growing similarly can differ by a factor of two when the mean serial interval differs by a factor of two. This shows that estimates of *R* obtained by assuming a form for the serial interval distribution come with the risk of substantial estimation bias.

The *basic* reproduction number (R_0) is the mean number of infections generated by a 'typical' infective person in a community with everyone susceptible and no public health interventions in place. Throughout this paper, *R*

refers to the *initial* reproduction number. For pandemic influenza this is likely to differ from R_0 for two reasons. Firstly, some cross-immunity from previous exposure to influenza strains may be present. Secondly, prior alertness to the possibility that the pandemic strain may be imported, and its unknown severity, may result in atypical behaviour and an enhanced public health response.

Wallinga and Teunis [5] provide a method for estimating R that is based on considering, for every case, who might have been responsible for that infection. The distribution of the serial interval is assumed to be known. Cauchemez *et al*. [6] modified the approach to enable dynamic estimation of R over time. The above comparison for case series A and B suggests that it is preferable to estimate R and the mean serial interval simultaneously from early data. A method for making Bayesian inferences about R, without assuming a specific distribution for the serial interval, is proposed by Cauchemez *et al*. [7]. They assume that a certain fraction of infections are traced as an epidemic progresses. An alternative approach to estimating R during the early stage of an epidemic is described by White and Pagano [8]. Their results suggest that it is possible to estimate R and parameters of the serial interval distribution simultaneously using only daily incidence data. Inspection of case series A of Table 1 indeed suggests that there is scope to estimate R from incidence data alone when the pattern of incidences is strongly suggestive of the serial interval. When growth in incidence exhibits waves over time, we can regard the sources of infection in one wave to be the cases of the previous wave, as illustrated for smallpox by Becker [9]. Here we consider estimation of R by both maximum likelihood and Bayesian methods. The aim is to determine what data are needed to make the estimate of R precise enough to inform decisions on public health interventions.

Table 1

Daily incidence counts of four artificial incidence series

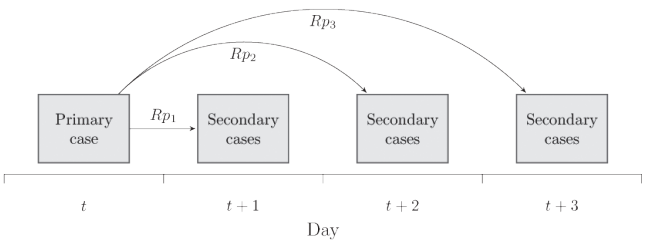
Incidence series	Daily incidence counts									
	Day[a]									
	-1	0	1	2	3	4	5	6	7	8
A	–	1	0	4	0	16	0	64	0	256
B	–	1	2	4	8	16	32	64	128	256
C	–	10	0	20	0	40	0	80	0	160
D	4	6	8	12	16	24	32	48	64	96

[a] Day 0 is the day initial cases present. In series D, four additional initial cases present on the previous day.

Four series of daily incidence counts that coincide with the mean count when $R = 4$ for series A, $R = 2$ for series B, C and D, and a short symptomatic infectious period on the first day after infection in series B and on the second day after infection in series A, C and D.

Figure 1

The infection process: mean number of secondary cases generated by a single case



p_i; probability that a serial interval is i days; R: initial reproduction number; t: day of symptom onset.

Note that Rp_i is simply R multiplied by p_i.

Mean number of secondary cases, with onset of symptoms in the next three days, generated by a single case with onset of symptoms on day t.

Methods

The alert of a possible pandemic virus strain instigates enhanced surveillance of incoming travellers and the general population. It is therefore possible to have daily incidence data of reasonable quality during the early stage of a detected outbreak. Observations on serial intervals are harder to collect because it is often difficult to ascertain the source of an infection. However, the first cases of a newly emerged infection are often travellers and subsequent local cases can sometimes be linked to incoming infected travellers. This can provide observations of serial intervals, as can sequential cases in early household outbreaks.

For maximum likelihood estimation and Bayesian inference of R, we need a likelihood function. We use the likelihood function proposed by White and Pagano [8], which is based on the infection process depicted in Figure 1, with one modification. We augment the likelihood with a contribution for independent observations of the serial interval, as described in the Appendix. We also use an unrestricted range of distributions for the serial interval, so we can better explore how results depend on the shape of this distribution. This likelihood function was used for maximum likelihood estimation and in Bayesian inferences via Markov chain Monte Carlo (MCMC) methods on simulated data [10], to see how these inferences perform with different amounts of data and with different rates of disease transmission and progression.

For our assessment of data needs we simulated, for each choice of parameter values, a large number of outbreaks, as in White and Pagano [8]. Specifically, we begin an outbreak with a fixed number of newly infected individuals. We assume that the number of infections generated by an infected case has a Poisson distribution. This assumes that each case has the same potential to infect others. We also assume that the serial interval has a multinomial distribution, as depicted in Figure 1.

We covered values of R in the range one to five, and a wide range of plausible shapes for the distribution of the serial interval.

How precise should the estimate of R be? We note that a precise estimate is valuable when R is near one, because it is then useful to be assured that a small amount of additional intervention, such as use of antiviral drugs or restricted school attendance, may contain the outbreak. When R is large (e.g. $R > 3$), we know that considerable intervention is required and the precise value of R is not quite as critical. On this basis, we aim for a precision so that the lower and upper values of a 95% credibility interval lie 25% or less below and above the value of R, respectively.

Results

As mentioned, data needs were investigated by assessing estimates obtained from many simulations of randomly generated outbreaks. Illustrative results for such simulated outbreaks are given for a few combinations of parameter values in Table 3 in the Appendix. This comprehensive assessment of the methods of inference from such simulated outbreaks led to several useful findings. Here we report these findings with reference to four simple illustrative incidence series, specifically chosen to point to the underlying reasons for the results.

First, we found circumstances when a useful estimate for R can be obtained from daily incidence data alone and having independent observations on the serial interval does not improve the precision of the estimate appreciably. This point is illustrated by estimating R from the case incidences shown in series A of Table 1. Let p_i denote the probability that a serial interval is i days. Incidences A coincide with the mean counts obtained from the model when $R = 4$ and the serial interval is two days (i.e. $p_2 = 1$). The mean serial interval (μ) is then two. Without additional observations on the serial interval, Bayesian inferences (described in the Appendix) gave the 95% credibility intervals for R, p_1, p_2, p_3 and μ shown in row one of Table 2. Note that:

- an R value of four lies in the 95% credibility interval for R and the interval bounds are only about 10% below and above four
- a μ value of two lies in the 95% credibility interval for μ
- the large value for p_2 and small values for p_1 and p_3 are indicated well by the inferences.

Specifically, note the tight credibility interval for μ, although no independent observations on the serial interval are included.

The above illustration is for incidences artificially chosen to coincide with the mean incidence counts, when $R = 4$ and $p_2 = 1$. Similar performance was observed when incidences were simulated to include a chance component (see Table 3 in the Appendix, for an illustration). The conclusion that incidence data alone can provide useful estimates also holds for variable serial intervals. This is illustrated by results in Table II of White and Pagano [8], who assume certain gamma distributions for the serial interval.

Row two of Table 2 shows the credibility intervals obtained when, in addition, there are 20 observations on serial intervals consisting of 18 serial intervals of two days and one serial interval of each of one day and three days. It is seen that adding the independent observations on serial intervals does not improve the precision of inferences. A similar conclusion is reached from the properties of maximum likelihood estimates. Specifically, the large sample standard deviation of the maximum likelihood estimator for R, with parameter values as for series A, is the same (to four decimal places) whether the number of observations on the serial interval is zero or 20.

The extreme pattern of incidences in A is very suggestive of a mean serial interval of two. More generally, we found that incidence data alone provide a good estimate whenever the serial interval distribution is unimodal and the mode is greater than one day. With such a serial interval distribution, a wave pattern tends to be superimposed on the exponentially growing incidence counts, and this pattern is informative about the mean serial interval. In particular, the four gamma

Table 2

95% credibility intervals from the daily incidence counts of four artificial incidence series[a], with and without additional observations on serial intervals

Row	Incidence series	95% credibility intervals of the parameter				
		R	p_1	p_2	p_3	μ
1	A	3.64–4.46	0.00–0.01	0.96–1.00	0.00–0.04	2.00–2.04
2	A + 20[b]	3.66–4.52	0.00–0.02	0.94–0.99	0.00–0.05	1.99–2.05
3	B	2.11–4.88	0.10–0.88	0.01–0.70	0.01–0.71	1.16–2.51
4	B + 20[b]	1.97–2.59	0.67–0.95	0.01–0.23	0.01–0.21	1.07–1.52
5	C	1.81–2.25	0.00–0.01	0.97–1.00	0.00–0.03	1.99–2.02
6	D	1.40–2.06	0.27–0.91	0.01–0.55	0.01–0.51	1.12–2.14
7	D + 10[b]	1.56–2.14	0.15–0.56	0.31–0.76	0.02–0.31	1.50–2.08

μ: mean serial interval; p_i: probability that a serial interval is i days; R: initial reproduction number.

[a] The four artificial incidence series (A–D) in Table 1.

[b] The number after the plus sign is the number of observations on the serial intervals.

distributions used by White and Pagano [8] are unimodal and have a mode greater than one day, which is what enables the incidence data alone to produce useful estimates.

In contrast, we also found circumstances when observations on daily incidence data alone are inadequate for simultaneous estimation of R and parameters of the serial interval distribution. We illustrate this observation by estimating R from the case incidences shown in series B of Table 1.

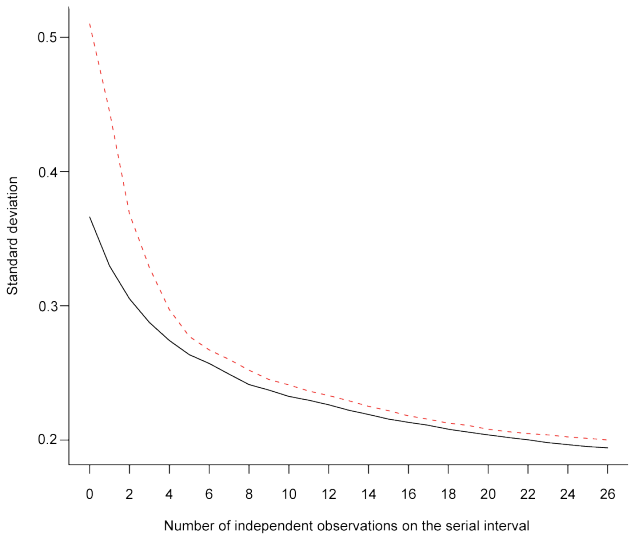
The daily incidences in series B coincide exactly with the mean incidence counts when $R = 2$ and $p_1 = 1$, so that $\mu = 1$. These parameter values are not recovered well by Bayesian estimation applied to the incidence data alone, as shown by the credibility intervals in row three of Table 2. The interval for R is wide and does not contain the value $R = 2$. Inferences about the distribution of the serial interval do not suggest a value near one for p_2, nor for μ. The four gamma distributions assumed for the serial interval by White and Pagano [8] do not reveal this weakness in making inferences from incidence data alone. By adding 20 independent observations on serial intervals (18 serial intervals of one day, one of two days and one of three days), the width of the credibility intervals narrows appreciably (see row four of Table 2). The main reason for the poor inference when there are no observations on serial intervals lies in the fact that the growing incidence in series B displays no wave pattern, so the incidence data provide minimal information about the mean serial interval. More generally, we found that the precision of estimates of R from incidence data alone is poor when the probability that serial interval is less than two days exceeds 0.5. Specifically, with a gamma distribution for the serial interval (as in [8]), estimation is poor when the mode of the distribution is zero, e.g. the exponential distribution. In such instances, estimation improves substantially by adding observations on the serial interval.

The following is a useful warning about choosing a suitable value for the number of initial infected individuals in simulation studies that assess methods for estimating R. It is natural to avoid very small outbreaks in simulation studies because they provide little information for estimation and in practice are unlikely to lead to attempts to estimate R. It is therefore common practice to start a simulated transmission chain with a larger number of initial cases. For example, White and Pagano [8] and Cauchemez *et al.* [6] generally start with 10 initial cases, and sometimes with 100. We found that assessing inferences based on 10 cases on the initial day tends to suggest better precision than is likely with more realistic initial case clusters. We illustrate this point by comparing inferences for case series C and D of Table 1. Incidence series C begins with 10 cases on day 0, while incidence series D begins with six cases on day 0 and four cases with onset of symptoms the previous day. With those respective initial cases, incidence series C and D coincide exactly with the mean counts when $R = 2$ and $p_2 = 1$. Both series have the same number of cases over the 10-day observation period, so the two series might be expected to contain approximately the same amount of information. The credibility intervals shown in row 5 and row 6 in Table 2 show that inferences for incidence series C are more precise than those for series D. Incidences C lead to better precision, particularly for p_1, p_2, p_3 and μ, because the 10 initial cases generate a better wave pattern on the exponentially growing incidences than does the initial case cluster in series D.

Given that incidence data alone are insufficient for estimating R for all plausible incidence series, it is important to determine how many observations on serial intervals are necessary to estimate R, when incidence data of an outbreak are inadequate for such estimation. Analyses based on Bayesian inferences and on maximum likelihood estimation indicate that just a few observations lead to a substantial improvement in the precision of estimates. This is illustrated in Figure 2. The solid curve shows the large-sample standard error of the maximum likelihood estimate of R as the number of observations on the serial interval increases. For this curve, we started with 10 cases on the first day and observed the incidence over the following four days, assuming $R = 2$ and that the serial interval has the distribution given by $p_1 = 0.61$, $p_2 = 0.32$ and $p_3 = 0.04$ (a distribution of the binomial form). The dashed curve shows the standard deviation for the posterior distribution of R when we have two initial infective cases and incidence counts of 2, 4, 7, 12, 21, 36, 61 and 104 over the next eight days. These counts are the mean counts (rounded to the nearest integer) when

Figure 2

Effect of increasing the number of observations on the serial interval



Large-sample standard deviation of the maximum likelihood estimate (solid line) and standard deviation of the posterior distribution (dashed line) of the initial reproduction number (R) as the number of observations on the serial interval increases.

$R = 2$ and the serial interval distribution is given by $p_1 = 0.61$, $p_2 = 0.32$ and $p_3 = 0.04$. Both curves illustrate the important point that the first few observations improve precision substantially. After 10–15 observations, each additional observation provides only a modest improvement in precision. This is typical of other settings where the incidence data alone are inadequate for estimating R with a precision of practical value. Note that both curves in Figure 2 decrease to a positive value. This value depends on the amount of incidence data available, which constrains the precision that is possible when estimating R.

It remains to ask how long a series of incidence data needs to be observed before we can estimate R with useful precision. This depends on the value of R and on the distribution of the serial interval. However, useful guidance is found by noting that estimating R corresponds to estimating the mean of the 'offspring' distribution, and so the number of infective individuals who are 'parents' is key to answering that question. As generations are not identified, some idea about the mean serial interval is needed. We found that R can be estimated with useful precision if we wait until the cumulative incidence reaches 150 and then continue to observe incidence for a number of days equal to the mean serial interval. Then the incidence data will include close to 150 parents (primary cases). This is illustrated by the results in row 5 (series C) in Table 2 when the incidence data are informative about the serial interval and by the results in row 6 (series D+10) when the incidence data contain little information about the serial interval. Note that series C and D each include 150 parents (primary cases) and there are 160 cases in the final generation whose offspring have not yet been observed.

Discussion

For a disease such as severe acute respiratory syndrome (SARS), with a latent period of a few days and onset of symptoms at about the start of the infectious period, it is very likely that the modal value of the serial interval is located a few days past the point of infection. Our results indicate that R can then be estimated quite effectively from daily incidence data alone. In contrast, for influenza the latent period and time to onset of symptoms tends to be quite short and individuals are thought to be infectious prior to onset of symptoms. It is not clear that the serial interval for the next influenza strain will have a modal value greater than one day. It is therefore sensible to include plans for observing some serial intervals into preparedness plans for pandemic influenza. As few as 10 observations can improve the precision of the early estimate of R substantially. The point that serial interval data improve the estimation of R was also made in the recent paper by White *et al*. [11].

In contrast to the approach of Cauchemez *et al*. [7], our approach assumes we have *independent* observations on the serial interval. This assumption made it feasible for us to carry out the analysis for many choices of parameter values. The assumption has no impact on results for the performance of estimates without serial interval data. For results that include serial interval data, we note that the independence assumption holds when the serial interval data come from a different location. When the serial interval observations are part of the locally collected incidence data, there is some dependence that is ignored by our approach. This is unlikely to have a significant impact on results, since we are assuming we have serial interval observations for less than 10% of infections.

The difficulty of observing serial intervals is exacerbated by the fact that the serial intervals actually observed may not be truly representative of randomly selected serial intervals, because they often arise from household contacts (with higher rates of contact within households) and from infected travellers (who may not have spent all of their infectious period locally).

It is important to be aware that biases may arise from the use of early incidence counts. First, it is important to allow for imported infections. Each imported case must be considered to have been infected elsewhere and not an offspring of a case from an earlier day. The methods used here are easily adapted to allow for this. Second, if a newly emerged infection is not detected immediately there may be a build-up of cases who are then detected in quick succession. Such a burst in the number of detected cases may not reflect the natural history of the infection and disease progression and can lead to initial estimates being biased.

As mentioned previously, the inferences reported here assume that the number of infections generated by an infected case has a Poisson distribution. This assumes that each case has the same potential to infect others and does not allow for variation in infectivity between individuals.

Our assessment clearly involved assumptions. During the early stages of the next newly emerged pandemic strain, it will not be known how appropriate these assumptions are. It will nevertheless be very useful to use these results on the type and quantity of data needed for guidance in preparedness plans for future emerging infections.

Appendix:
http://nceph.anu.edu.au/Staff_Students/Staff%20Publications/Appendix_Becker.pdf

Acknowledgements

This work was supported by Australian National Health and Medical Research Council (NHMRC) grants 471436 and 585536.

References

1. Osterholm MT. Preparing for the next pandemic. N Engl J Med. 2005;352(18):1839-42.
2. World Health Organization (WHO). Avian influenza. Geneva: WHO. Available from: http://www.who.int/csr/disease/avian_influenza/en/
3. Enserink M. H5N1 vaccine stockpile plan advances. ScienceNOW. 25 April 2007. Available from: http://news.sciencemag.org/sciencenow/2007/04/25-03.html
4. H5N1 pre-pandemic vaccine plans in Japan. Recombinomics. 18 April 2008. Available from: http://www.recombinomics.com/News/04180803/H5N1_Pre_Pandemic_Japan.html
5. 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.
6. Cauchemez S, Boëlle P-Y, 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.
7. Cauchemez S, Boëlle P-Y, Thomas G, Valleron A-J. Estimation in real time the efficacy of measures to control emerging communicable diseases. Am J Epidemiol. 2006;164(6):591-7.
8. White LF, Pagano M. A likelihood based method for real time estimation of the serial interval and reproductive number of an epidemic. Stat Med. 2008;27(16):2999-3016.
9. Becker N. Estimation for an epidemic model. Biometrics. 1976;32(4):769-777.
10. Sorensen D, Gianola D. Likelihood, Bayesian and MCMC methods in quantitative genetics. New York: Springer; 2002.
11. White LF, Wallinga J, Finelli L, Reed C, Riley S, Lipsitch M, et al. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza A/H1N1 pandemic in the USA. Influenza Other Respi Viruses. 2009;3(6):267-76.