Estimating reproduction numbers for adults and children from case data

K. Glass1,*, G. N. Mercer1, H. Nishiura2,3, E. S. McBryde4 and N. G. Becker1

1National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia
2PRESTO, Japan Science and Technology Agency, Saitama, Japan
3Theoretical Epidemiology, University of Utrecht, Utrecht, The Netherlands
4Department of Medicine, University of Melbourne, Melbourne, Australia

We present a method for estimating reproduction numbers for adults and children from daily onset data, using pandemic influenza A(H1N1) data as a case study. We investigate the impact of different underlying transmission assumptions on our estimates, and identify that asymmetric reproduction matrices are often appropriate. Under-reporting of cases can bias estimates of the reproduction numbers if reporting rates are not equal across the two age groups. However, we demonstrate that the estimate of the higher reproduction number is robust to disproportionate data-thinning. Applying the method to 2009 pandemic influenza H1N1 data from Japan, we demonstrate that the reproduction number for children was considerably higher than that of adults, and that our estimates are insensitive to our choice of reproduction matrix.

Keywords: influenza; reproduction number; type structure; outbreak data

1. INTRODUCTION

When assessing how fast an infectious disease is spreading, there are two key disease characteristics that we need to know: the reproduction number and the serial interval. The effective reproduction number tells us how many people (on average), a randomly selected infected person infects. It differs from the basic reproduction number in that it incorporates the effects of susceptible depletion and any public health interventions, and so is more closely tied to empirically observed data. The mean serial interval is the average delay between one person showing symptoms and the individuals they infect showing symptoms. While the reproduction number may vary by location, over time and according to the level of intervention, we generally expect the serial interval to be consistent between locations and through time, unless rigorous extrinsic measures (such as case isolation and quarantine) influence the contact patterns [1]. In the event of an outbreak of an existing disease, while the distribution of the serial interval may be assumed to be known, dynamic estimates of the reproduction number are urgently needed for the policy makers to assess the local spread of the disease and to determine the probable impact of interventions.

In recent years, a number of methods have been developed for estimating the reproduction number from outbreak data [2–10]. Some of these methods estimate characteristics of the serial interval simultaneously [2,4,9], while others assume that the distribution of the serial interval is known [5–8]. Most methods can only be applied during the early ‘take-off’ phase of an outbreak, and so provide a single estimate of the reproduction number, but some—such as the Wallinga & Teunis method [7] with Cauchemez adjustment [3] (similar to that used in Ferguson et al. [11])—estimate the time-varying reproduction number over the course of the outbreak.

While necessary for assessing overall disease spread, estimates of a single reproduction number for an entire population do not provide information on the underlying heterogeneous transmission within and between different types of individuals. For example, some diseases show an age-specific pattern of infection, with a higher number of cases in a particular age group. Alternatively, some diseases show higher case numbers in a particular ethnic group or other sub-population. Both of these situations occurred during the 2009 pandemic influenza H1N1 outbreak [12,13]. Type structure is routinely included in disease transmission models to allow for dependence on age or other forms of population structure [14,15]. Here, we estimate type-specific reproduction numbers, using adults and children as a case study. This approach provides estimates of the average number of secondary cases that a single individual of each type produces. A comparison of these type-specific reproduction numbers provides us with an understanding of how much adults and children contribute to transmission, which can inform decisions on interventions.

Recent work by Nishiura et al. [16] incorporates age structure into estimates of the reproduction

* Author for correspondence (kathryn.glass@anu.edu.au).
number for a single time period. Here, we present methods for estimating reproduction numbers for two types of individuals over time, allowing us to detect the impact of interventions on different groups, and to monitor changes over the course of an outbreak. While the primary method presented here is an extension of the transmission network method used by Wallinga & Teunis [7], we also compare this with the population growth method used by White & Pagano [9,10].

For simplicity, we will refer to these approaches as the ‘Wallinga & Teunis’ [7] and the ‘White & Pagano’ methods [9,10], although we acknowledge that others have used similar techniques, as outlined above. In this paper, our aim is to quantify the type-specific transmission patterns using daily case numbers stratified by type. We also assess the sensitivity of these estimates to assumptions about transmission between the groups and to under-reporting of case numbers. These methods are applied to 2009 influenza pandemic H1N1 (pH1N1) data from Japan to compare reproduction numbers for adults and children.

2. METHODS

2.1. Data

We use both computer-generated simulation data and real pH1N1 incidence counts to test the methods. We do not attempt to estimate the reproduction numbers and characteristics of the serial interval simultaneously, but adopt the distribution of the serial interval estimated from Australian data at the start of the outbreak [17] for both types of individuals. This has a gamma distribution with mean of 2.9 days and s.d. of 1.4 days.

Simulated data are generated from a stochastic transmission model using the various reproduction matrices outlined below. In each simulation, we assume one initial case of type selected according to the proportion in the population. Simulations are assumed to take off if 50 or more cases occur, and are run until there are 2000 cases. In general, 100 simulations are conducted for each scenario. We tested the impact of the type of initial case (i.e. either adult or child) on estimates of the type-specific reproduction numbers and found only an early, short-term effect and no effect on the final estimates. The short-term impact of the initial case is no longer visible once there are around 20 cumulative cases of either type, which is the time at which we begin plotting the time-varying reproduction numbers. To test the effect of under-reporting, we consider outbreaks in which a fraction of either adult cases, child cases or both are not reported. We also apply the method to pH1N1 data from Japan described in the study of Nishiura et al. [16].

2.2. Reproduction matrices

The reproduction matrix for two types of individual has the form

\[
M = \begin{pmatrix}
    m_{CC} & m_{CA} \\
    m_{AC} & m_{AA}
\end{pmatrix},
\]

where \(m_{ij}\) denotes the average number of cases of type \(i\) infected by a single individual of type \(j\) throughout its entire course of infection. In a fully susceptible population, this matrix is often referred to as a next-generation matrix [18].

A key issue is that we can estimate only two parameters for the \(2 \times 2\) matrices, as daily incidence counts give us only two growth rates. If we attempt to estimate all four elements of this matrix from daily case data, we generally find a band of equal-likelihood values pairing matrix elements \(m_{CC}\) with \(m_{CA}\) and \(m_{AC}\) with \(m_{AA}\) (figure 1). In order to estimate two reproduction numbers unambiguously, we must make some assumptions about the form of the reproduction matrix. Here, we consider the influence of the matrix parametrization on the resulting interpretation of the transmission dynamics for the four matrix forms.

2.2.1. Symmetric matrices

In analysing the early pH1N1 Japanese data, Nishiura et al. [16] consider two forms for the reproduction matrix: a separable matrix and one with high child-to-child transmission (which we will refer to as HiC2C). The separable reproduction matrix can be separated into row and column parameters, while the HiC2C matrix assumes a high transmission rate between children, with all other parameters equal. This is one of the primary WAIFW matrices adopted by Anderson & May [14] for two age groups. See table 1 for the formulation of these matrices.

Within the class of symmetric reproduction matrices, these two forms are the ones most suited to estimating reproduction numbers for two types. Other forms of reproduction matrices that are type-specific or which assume that infectivity or susceptibility vary with type are either too restrictive, or cannot be estimated from daily incidence data. See appendix A for further details on this.

2.2.2. Asymmetric matrices

Symmetric matrices are not suitable in all cases. In the case of adults and children, while all children have contact with adults, many adults in the population have little contact with children. That is, we expect the element of the matrix indicating adult–child infection to be smaller than the child–adult element. In the case of the first wave of transmission in Japan [16], the outbreak was centred on a school, so a symmetric pattern of infection is not unreasonable for these initial cases. When the outbreak involves wider community transmission, the assumption of symmetry is less likely to be valid, as many adults involved in the outbreak will not have regular contact with children.

For cases where symmetric reproduction matrices are not appropriate, there are many types of matrix form we could use. Here, we consider two asymmetric matrices: a contact-frequency matrix that uses data on contact patterns to determine mixing between the two groups, and a modified proportional mixing matrix that uses data on the proportion of the population of each type to determine transmission patterns (table 1). It has been recognized that human contact networks tend to be highly assortative—that is, individuals prefer to associate with individuals of the same
— and both of these matrix forms allow for assortative mixing.

The contact-frequency matrix is applicable in the case where there are data on the mixing patterns of the two groups. For adults and children, we use data presented in Mossong et al. [20] to calibrate mixing patterns, and assume that 50 per cent of a child’s contacts are with children, while 75 per cent of an adult’s contacts are with other adults (that is $r_C = 0.5$ and $r_A = 0.75$). Under this assumption, it should be noted that we do not expect the matrix to satisfy reciprocity [21]; our contact-frequency matrix also includes two parameters to be estimated that allow for differences in infectivity and susceptibility between the two types of individual.

The modified proportional matrix assumes that adults infect other individuals according to their proportions in the population. We assume that children make up 25 per cent of the population, so that 25 per cent of an adult’s infections are children and 75 per cent are other adults. We then adjust this proportional matrix to allow for extra transmission between children. For simplicity, we will refer to this matrix form as the proportional reproduction matrix.

Finally, we also consider matrices that do not correspond to any of the above formats to assess the effect of the imposed structure on the estimated reproduction numbers.

### 2.3. Definition of the type-specific reproduction number

We define the reproduction number for children (adults) as the average number of cases generated by a single infected child (adult) at calendar time $t$. This quantity is calculated in a different manner in the following two algorithms, but both methods use daily incidence data to fit a transmission matrix of the form $M = [m_{ij}]$. The White & Pagano method [9,10] estimates single values for the components of the matrix $M$, assuming that this matrix remains constant over the initial stages of the outbreak. The Wallinga & Teunis method [7] reconstructs the transmission network in such a way that there is a daily estimate for the reproduction matrix $M$. For both approaches, the reproduction numbers for the two types of individual are given by the column sums of $M$, that is:

$R_C = m_{CC} + m_{AC}$ and $R_A = m_{CA} + m_{AA}$.

We use $R$ to refer to the dominant eigenvalue of the matrix $M$—that is, the population reproduction number. It should be noted that $R_C$ and $R_A$ are different from the so-called type-reproduction number.
proposed by Roberts & Heesterbeek [22] for computing required public health effort to eradicate an infectious disease. Rather, our $R_C$ and $R_A$ represent the average numbers of secondary cases (inclusive of both child and adult secondary cases) generated by a single child and adult primary case, respectively, and we hereafter refer to these as the ‘type-specific reproduction number’. These type-specific reproduction numbers are useful for assessing ‘I-control’ [23], where interventions act on the infectiousness of the target group, for example, through case isolation and social-distancing measures such as school closure. It is also possible to adapt the methods proposed here to sum row entries of $M$ to assess ‘S-control’, where interventions modify susceptibility of the target group—for instance, using vaccination or antiviral prophylaxis.

### 2.4. White and Pagano method for estimating type-specific reproduction numbers

Here, we adapt the method used by White & Pagano [9,10] to allow for two types of individual. In this study, we restrict the use of this method to an estimation of the reproduction number during the early ‘take-off’ period of the epidemic. That is, we assume that the transmission matrix is constant over this period, and estimate a single value for each of the two type-specific reproduction numbers. Assume that we have $T$ days of daily incidence data $C_t$ for children and $A_t$ for adults, and the serial interval has a distribution with a probability function $w(\tau)$, $\tau = 1, \ldots, k$. In a manner similar to White & Pagano [9,10], we define $X_i \sim \text{Pois}(m_{CC}C_i + m_{AC}A_i)$ and $Y_i \sim \text{Pois}(m_{CA}C_i + m_{AA}A_i)$ to be the total number of cases of each type infected by individuals with onset on day $i$. The derivation of the likelihood described in White & Pagano [9,10] can then be reproduced for two types, where we define $\mu_A$ and $\mu_C$, the expected number of adult and child cases on day $t$ as

$$ \mu_C(M, t|C_t, A_t, w) = m_{CC} \sum_{\tau=1}^{\min(k, t)} C_{t-\tau} w(\tau) $$

$$ + m_{CA} \sum_{\tau=1}^{\min(k, t)} A_{t-\tau} w(\tau) $$

and

$$ \mu_A(M, t|C_t, A_t, w) = m_{CA} \sum_{\tau=1}^{\min(k, t)} C_{t-\tau} w(\tau) $$

$$ + m_{AA} \sum_{\tau=1}^{\min(k, t)} A_{t-\tau} w(\tau). $$

The summation over $\tau$ includes all cases that may infect a new case on day $t$, adjusted by the probability of a serial interval of duration $\tau$. The likelihood function then has the form

$$ L(M|C_t, A_t, w) = \prod_{t=1}^{T} \frac{e^{\mu_C(t)C_t}}{C_t!} \prod_{t=1}^{T} \frac{e^{\mu_A(t)A_t}}{A_t!}. $$

It is possible to maximize this function with no restrictions on the matrix elements $m_{ij}$, however, it is often difficult to distinguish all terms. Figure 1 is an illustrative plot of the likelihood function for each of the six pairs of variables when estimating matrix elements from computer-generated data. There is a band of approximately equal likelihood for the pair $m_{CC}$ and $m_{CA}$ and for the pair $m_{AA}$ and $m_{AC}$. This pattern is typical for pH1N1 data as well as computer-generated data.

However, if we maximize the likelihood function subject to one of the reproduction matrix forms discussed earlier and given in table 1, the four matrix elements are reduced to two parameters in such a way that the identifiability problem is removed. It is then possible to derive unambiguous estimates for the reproduction matrix entries and the reproduction numbers. Provided the fraction of infected individuals of each type remains approximately constant, we can derive the following two equations in four unknowns that are satisfied by the maximum-likelihood solutions for all matrix forms.

$$ C = m_{CC}C + m_{CA}A $$

and

$$ A = m_{AC}C + m_{AA}A, $$

where

$$ C = \sum_{t=1}^{T} C_t, $$

$$ A = \sum_{t=1}^{T} A_t, $$

$$ \hat{C} = \sum_{t=1}^{T} C_t, $$

and

$$ \hat{A} = \sum_{t=1}^{T} A_t. $$

Note that if the fraction of infected individuals of each type varies over the course of the outbreak, the time-varying reproduction number produced by the Wallinga & Teunis method [7] (which allows for these changes) is more appropriate.

Together with a choice of matrix form, this will allow maximum-likelihood estimates to be calculated algebraically. For instance, with a separable matrix, where $m_{AC} = m_{CA}$ and $m_{AA}m_{CC} = m_{AC}m_{CA}$, we have:

$$ R_C = m_{CC} + m_{AC} = \frac{\hat{C} + \hat{A}}{C + A} $$

and

$$ R_A = m_{CA} + m_{AA} = \frac{\hat{A}}{C + A}. $$

Further details for separable and other matrix forms are given in appendix A.

### 2.5. Wallinga and Teunis method for estimating type-specific reproduction numbers

Here, we adapt the methods used by Wallinga & Teunis [7] and Cauchemez et al. [3] to allow for two types of individual. These methods for reconstructing the transmission network (or transmission tree) are similar to the estimation framework of the cohort reproduction number in mathematical demography, whose usefulness has recently become apparent in infectious disease epidemiology [24,25]. The Wallinga & Teunis method [7] provides daily estimates of the reproduction number, and thus can be applied to part or all of a disease outbreak. The algorithm estimates the probability that individual $i$ infected individual $j$ on the basis of the gap between the onset days of the two individuals.
Assume that case $i$ has onset time $t_i$ and is of type $a_i \in \{C,A\}$. Recall that $w(t), t = 1, \ldots, k$ is the distribution of the serial interval (common to both adults and children). For a single type of individual, Wallinga & Teunis [7] define

$$p_{ij} = \frac{w(t_j - t_i)}{\sum_{l \neq j} w(t_j - t_l)},$$

to be the probability that individual $i$ was infected by individual $j$. They then used this to estimate a reproduction number for individual $j$ as $\sum_{i \neq j} p_{ij}$.

We modify these formulae to take account of the type of individuals $i$ and $j$. Define a weight, $q_{ij}$, to be the fraction of child cases that are infected by a child (so that $1 - q_{ij}$ is the fraction of children that are infected by an adult). Similarly, $q_A$ is the fraction of adult cases that are infected by an adult. When assessing the probability that individual $j$ infected individual $i$, we then compare the gap between onset of $i$ and $j$ with that of other individuals of the same type as $j$. That is, we have

$$p_{ij} = \begin{cases} q_{ij} w(t_j - t_i) & \text{if } a_i = a_j, \\ \sum_{j, q_{ij}} w(t_j - t_i) & \text{if } a_i \neq a_j. \end{cases}$$

Note that the denominators of the type-specific estimates sum over individuals, $l$, of the same type as individual $j$ only. This is done to ensure that the probability that individual $i$ was infected by an individual of the same type is $q_{ij}$—that is, the probability that a child is infected by a child is $q_{ij}$ in line with the weights defined above. This can be verified algebraically as $\sum_{j, q_{ij}} p_{ij} = q_{ij}$ and $\sum_{j, q_{ij}, a_i} p_{ij} = 1 - q_{ij}$. As before, the reproduction number for individual $j$ is given by $\sum_{i \neq j} p_{ij}$.

We then calculate daily reproduction numbers for the two types of individuals as

$$R_C(t) = \sum_{i,j : a_i = C} p_{ij} \quad \text{and} \quad R_A(t) = \sum_{i,j : a_i = A} p_{ij},$$

that is, the sum of the estimated number of cases infected by individuals of that type with onset on day $t$.

It remains to estimate $q_C$ and $q_A$ given the daily incidence counts and a choice of reproduction matrix form. Define $f$ to be the fraction of total cases that are children. Then, for any of the reproduction matrices in table 1, it is possible to express the matrix elements (and thus $q_C$ and $q_A$) in terms of $f$ and the population reproduction number, $R$. For instance, for a separable matrix, we can calculate $q_C = 1/(1 + x^2)$ and $q_A = x^2/(1 + x^2)$, where $x = (1 - f)/f$.

During the initial exponential growth phase of an outbreak, we would expect $f$ to be relatively constant, but often $f$ will change over the course of the outbreak. In this case, we use daily values of $f$ to calculate $q_C$ and $q_A$, and thus allow them to vary with time. Details of $q_C$ and $q_A$ for various matrix forms are given in appendix A.

### 2.6. Effect of thinning

It is rare that every infected individual is included in surveillance data, particularly if the disease is common or mild. Analysis of methods for estimating the reproduction number for a single type of individual shows that they are robust to under-reporting, even if only a very small proportion of cases are reported [26], provided the proportion of cases reported remains constant over time [27]. When a change in the reporting proportion occurs, this generally impacts the time-dependent estimates of the reproduction number for a time period approximately equal to twice the mean serial interval [28]. For a disease such as influenza, such effects are relatively minor provided they do not occur too frequently. When we consider the impact of thinning on estimates of reproduction numbers for adults and children, we find that they are similarly robust if the proportion of cases reported is equal across the two age groups. However, if the extent of thinning varies between the groups—say because health-workers have been instructed to prioritize the follow-up of at-risk groups—there will be some effect on the estimates of the reproduction numbers.

We investigate the effect of thinning analytically and using simulated datasets with different reproduction matrices assuming that the thinning is constant through time. For a matrix form, we can define the reproduction numbers of the two types in terms of the local population reproduction number ($R$) and the fraction of cases that are children ($f$). As discussed above, $R$ is relatively insensitive to thinning, so we focus on assessing the sensitivity of the type-specific reproduction numbers to inaccuracies in $f$.

### 3. RESULTS

#### 3.1. Simulation data

We begin by considering estimates of type-specific reproduction numbers from computer-generated data. In figure 2, we present estimates of these reproduction numbers using our modification of the Wallinga & Teunis method [7] from data generated using separable and proportional reproduction matrices for adults and children, assuming children make up 25 per cent of the underlying population. Figure 2a,c,e shows the analysis of the data generated using the separable matrix, and figure 2b,d,f shows the analysis using the proportional matrix. In each case, the matrices used in the simulations have a reproduction number for children ($R_C$) of 2.5 and a reproduction number for adults ($R_A$) of 1.0. Estimates of the reproduction number are shown as soon as the cumulative cases (across both groups) reaches 20. Figure 2a,b shows a typical example of simulated data. We see that the fraction of child cases is greater under the separable matrix when compared with the proportional matrix (71% children with the separable matrix versus 60% children with the proportional matrix). As shown by figure 2c,d, the method provides reliable estimates of the reproduction numbers when the same form of reproduction matrix that generated the data are used. Figure 2e,f shows the effect of...
using a different matrix form. When the data are generated using the separable matrix (figure 2a,c,e), all methods estimate the higher reproduction number reasonably well, but the asymmetric matrix methods (contact-frequency and proportional) underestimate the lower reproduction number. When the data are generated using the proportional matrix (figure 2b,d,f), both symmetric matrices (separable and HiC2C) underestimate the higher reproduction number and overestimate the lower reproduction number.

We also perform these calculations using the modified White & Pagano method [9,10] that gives a single estimate and not a time-varying one. Using the correct matrices, we estimate \( R_C = 2.47 \), \( R_A = 1.02 \) for the separable matrix, and \( R_C = 2.63 \), \( R_A = 0.91 \) for the proportional matrix. Again, estimates of \( R_C \) are too low and \( R_A \) are too high if symmetric matrices are assumed for the proportional data, and estimates of \( R_A \) are too low if asymmetric matrices are assumed for separable data.

Table 2 explores this in more detail, giving the 3rd percentile, the median and the 98th percentile of estimates of \( R_C \) and \( R_A \) for 100 simulations of each type of reproduction matrix, where the values used to generate the data are \( R_C = 2.5 \) and \( R_A = 1.0 \). By giving lower and upper percentiles, we describe a range that contains 95 per cent of the simulations. We see that all methods perform well if the assumed matrix form matches the original matrix form (the upper diagonal blocks in the table). Assuming a symmetric reproduction matrix, when the original matrix is asymmetric gives poor estimates of both reproduction numbers, with \( R_C \) underestimated and \( R_A \) overestimated. Estimating an asymmetric reproduction matrix when the original matrix is symmetric leads to an underestimate of \( R_A \), but only a slight overestimate of \( R_C \). Using the wrong asymmetric matrix leads to errors in the estimate of \( R_A \), while both the contact-frequency and proportional reproduction matrices provide good estimates of either type of asymmetric data.

Finally, the last four rows in the table present estimates for simulations run with two unstructured matrices with \( R_C = 2.5 \) and \( R_A = 1.0 \). These two unstructured matrices are:

\[
\begin{bmatrix}
1.8 & 0.2 \\
0.7 & 0.8
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
2.0 & 0.75 \\
0.5 & 0.25
\end{bmatrix}
\]

and were chosen to ensure that the force of infection on children is higher than on adults, but such that the matrices do not fit any of the four matrix forms described in this paper. The first matrix was selected to ensure that \( m_{CA} > m_{AC} \), with the second (less realistic) matrix satisfying \( m_{AC} > m_{CA} \). We see that the methods assuming an asymmetric matrix perform better in...
estimating parameters from the first matrix, while the methods assuming a symmetric matrix perform slightly better on the second matrix. In both cases, the estimates of $R_C$ from the asymmetric method show little bias, although the estimate of $R_A$ from the second (less realistic) matrix is underestimated. Both of these unstructured matrices assume a higher force of infection on children, leading to higher attack rates in children relative to adults. While there may be scenarios in which the attack rate is greater in adults despite a higher reproduction number in children, we are unlikely to attempt to analyse such data with a method that tests for higher transmission in children.

### 3.2. Japanese data

Next, we apply these methods to data from the initial stages of the pH1N1 outbreak in Japan [16], which had extremely high case numbers in children. Figure 3 presents the estimates of the reproduction numbers for

![Graph](https://via.placeholder.com/150)

Figure 3. Estimates of the reproduction numbers for the Japanese outbreak in school children reported in the study of Nishiura et al. [16]. (a) The estimates of the reproduction number for children $R_C$ in black and the reproduction number for adults $R_A$ in grey using a separable matrix and showing 95% confidence intervals. Estimates begin when there are 20 cumulative cases, with a delay in the adult reproduction number because adult cases did not occur until the 13 May. (b) The graph compares the estimates for the separable (solid lines), HiC2C (dashed lines), contact-frequency (dotted-dashed lines) and proportional (dotted lines) matrices.

<table>
<thead>
<tr>
<th>assumed true</th>
<th>separable</th>
<th>HiC2C</th>
<th>contact-frequency</th>
<th>proportional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
<td>M</td>
<td>U</td>
<td>L</td>
</tr>
<tr>
<td>separable</td>
<td>$R_C$</td>
<td>2.32</td>
<td>2.50</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td>$R_A$</td>
<td>0.90</td>
<td>1.00</td>
<td>1.14</td>
</tr>
<tr>
<td>HiC2C</td>
<td>$R_C$</td>
<td>2.37</td>
<td>2.57</td>
<td>2.74</td>
</tr>
<tr>
<td></td>
<td>$R_A$</td>
<td>0.67</td>
<td>0.77</td>
<td>0.92</td>
</tr>
<tr>
<td>contact</td>
<td>$R_C$</td>
<td>2.04</td>
<td>2.26</td>
<td>2.42</td>
</tr>
<tr>
<td></td>
<td>$R_A$</td>
<td>1.22</td>
<td>1.40</td>
<td>1.54</td>
</tr>
<tr>
<td>proportional</td>
<td>$R_C$</td>
<td>1.96</td>
<td>2.20</td>
<td>2.38</td>
</tr>
<tr>
<td></td>
<td>$R_A$</td>
<td>1.29</td>
<td>1.44</td>
<td>1.63</td>
</tr>
<tr>
<td>unstructured 1</td>
<td>$R_C$</td>
<td>2.00</td>
<td>2.20</td>
<td>2.37</td>
</tr>
<tr>
<td></td>
<td>$R_A$</td>
<td>1.25</td>
<td>1.41</td>
<td>1.62</td>
</tr>
<tr>
<td>unstructured 2</td>
<td>$R_C$</td>
<td>2.41</td>
<td>2.54</td>
<td>2.70</td>
</tr>
<tr>
<td></td>
<td>$R_A$</td>
<td>0.61</td>
<td>0.69</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 2. Estimates of $R_C$ and $R_A$ for computer data generated under four different structured reproduction matrices and two unstructured matrices, all with simulated values of $R_C = 2.5$ and $R_A = 1.0$. The row indicates the matrix form used to generate the data and the column indicates the assumption made by the estimation method. For each row and column pair, 100 simulations were run and the 3rd (L), the median (M) and the 98th (U) estimates are shown for each of $R_C$ and $R_A$, thus providing a range that contains 95% of the simulations.
adults and children using the modified Wallinga & Teunis method [7]. For all matrix forms, the reproduction number for children is considerably higher than that for adults at the start of the epidemic, and remains higher for most of the outbreak. Moreover, although there is slight variability between the estimates of the reproduction numbers for adults with different matrices, the estimates of the reproduction number for children are very similar across all four reproduction matrices. Over the first 8 days of data, the mean estimates of the child reproduction numbers are all close to 2.8, while the reproduction numbers for adults ranges between 0.3 and 0.7.

Estimating the reproduction matrices for the first 8 days of data (corresponding to 156 cases) using the White & Pagano method [9,10] gives the following matrix estimates:

\[
M_S = \begin{bmatrix}
3.20 & 0.31 \\
0.31 & 0.03 
\end{bmatrix},
M_H = \begin{bmatrix}
3.20 & 0.29 \\
0.29 & 0.29 
\end{bmatrix},
M_C = \begin{bmatrix}
3.21 & 0.16 \\
0.31 & 0.05 
\end{bmatrix}
\]

and

\[
M_P = \begin{bmatrix}
3.29 & 0.09 \\
0.28 & 0.28 
\end{bmatrix}
\]

for separable (S), HiC2C (H), contact-frequency (C) and proportional (P), respectively. This corresponds to reproduction numbers for children of 3.51, 3.49, 3.52 and 3.57 and reproduction numbers for adults of 0.34, 0.58, 0.21 and 0.37. Again, the results are very consistent across matrix forms for children and show some variation for adults.

Overall, these estimates are a little higher than those of Nishiura et al. [16], but show a similar pattern of high transmission between children. The differences between our estimates and those in Nishiura et al. [16], are largely owing to the difference in the choice of serial interval—Nishiura et al. [16], assume a mean serial interval of 2 days, in contrast to the 2.9 days assumed here. If we adopt a serial interval with mean of 2 days, the estimates of the reproduction numbers for the separable matrix drop to \(R_C = 2.96, R_A = 0.29\) and the HiC2C matrix drop to \(R_C = 2.93, R_A = 0.47\). These are comparable, although slightly lower than the estimates in Nishiura et al. [16]. Throughout this work, we have assumed that the distribution of the serial interval remains constant throughout the epidemic, and is the same for adults and children. Should data be available to identify changes in the serial interval over time [29] or differences between adults and children, this could be incorporated into the estimates of the reproduction numbers.

### 3.3. Thinned data

In figure 4, we show the results of testing the effect of data thinning on the estimates of the reproduction numbers for two types of individual using computer-generated data with reproduction numbers \(R_C = 2.5\) and \(R_A = 1\). We perform 100 simulations for each of the four matrix forms, taking the final estimate of the reproduction numbers calculated by the method. The results are broadly consistent across matrix forms. If the data are thinned evenly across the two types, both estimates are unbiased. If the child data are thinned, the reproduction number for adults is overestimated and the reproduction number for children is slightly underestimated for some matrices. If the adult data are thinned, the reproduction number for adults is underestimated and the reproduction number for children is slightly underestimated.

We are also interested to see how these results compare for different underlying values of \(R_C\) and \(R_A\). We consider only the case where children are disproportionately represented in the data—that is, where \(R_C > R_A\), and the data are not disproportionately thinned to the extent that the proportion of child cases in the dataset is less than the proportion of children in the population. In figure 5, we repeat the analysis of figure 4 for the case where \(R_C = 2.5\) and \(R_A = 1.5\), where the data are thinned to 70 per cent of their original value. The results are similar to those in figure 4, although the estimate of \(R_C\) is more biased in the case where the child data are thinned, while the estimate of \(R_A\) is generally less biased when the adult data are thinned.

Analysis of the various matrix forms in appendix A confirms the results displayed above. Provided child cases are disproportionately represented, estimates of their reproduction number are robust to thinning, generally lying at most 20–30% away from the original value. In the case of the symmetric matrices, the condition that children be disproportionately represented means that over 50 per cent of cases must be in the high-transmission group. In the case of the proportional matrix, we require the fraction of cases to be greater than \(p\) (i.e. at least 25% of cases in children in our example). In the case of the contact-frequency matrix, we require \((1 - \rho_C)/(1 - f)(1 - \rho_A) > 1\), which translates to at least 33 per cent of cases in children.
In contrast, we have demonstrated that evidence of a higher reproduction number in children is robust to detecting a difference between the reproduction numbers. Additionally, if child case data are thinned much more extensively than adult data such that the resulting proportion of child cases in the dataset is roughly equal to the proportion of children in the population, we will not detect a difference between the reproduction numbers. In contrast, we have demonstrated that evidence of a higher reproduction number in children is robust to mixing and thinning assumptions. Moreover, by outlining four different mixing assumptions, we provide scope to compare the output under different reproduction matrices to assess the extent to which these assumptions may have influenced the results. This helps to identify the level of uncertainty with respect to contact patterns that exist for a particular dataset—as highlighted in Farrington et al. [30], mixing assumptions can have a profound impact on estimates of the reproduction number.

In many studies of infectious disease transmission, reproduction matrices are assumed to be symmetric. That is, they assume that if a randomly selected child infects $n$ adults on average, then a randomly selected adult will infect $n$ children on average. This is often not appropriate for mixing between adults and children, as many adults have very little contact with children. Adopting a symmetric reproduction matrix for data split into adults and children will only identify higher transmission from children if child cases outnumber adult cases. If children do not outnumber adults in the dataset, but are nonetheless disproportionately represented, an asymmetric matrix is generally more appropriate.

In this paper, we consider two asymmetric matrices in detail: a proportional matrix and a contact-frequency matrix. The proportional reproduction matrix allows for extra transmission between one group, with the remaining transmission determined by the proportion of each type of individual in the population. This is a relatively flexible matrix form, as data on the proportion of the population that is in a particular age or ethnic group are generally easy to obtain. The contact-frequency reproduction matrix uses data on the contact patterns of the two types of individuals to determine mixing patterns, and then uses data to estimate differences in the susceptibility or infectivity between the two types of individual. While in many cases this is a more realistic matrix form, it also relies on data that are often not available. We have calibrated our model using European data [20], however, we note that even within Europe there is some variability between countries. The true patterns for Japan—or any other country—may well differ from our assumptions. However, it is reassuring to observe that the estimates of the reproduction numbers derived from the proportional and the contact-frequency matrices are very similar, both for simulated and real data. In future work, we hope to explore further the possible matrix forms to identify more generally the type of information that can be estimated under different matrix assumptions.

When incidence data are collected for a widespread outbreak, we anticipate that they will be considerably thinned. That is, only a proportion of the total cases will be recorded in the dataset. When the reproduction number is estimated for the population as a whole using the Wallinga & Teunis method [7], the estimate is remarkably robust to thinning—provided the proportion of individuals recorded in the dataset does not change markedly over time [26]. If thinning is consistent across types, the same is true of the two-type estimates. If thinning differs between types, there can be biases in the estimates of the reproduction numbers, but this is most apparent for the type with lower reproduction number (adults in the examples considered here).
Provided the type with the higher reproduction number (children) are disproportionately represented in the data, estimates of this reproduction number have much less bias. Our simulations and analysis show that this arises because the reproduction number of the type with high transmission is tightly linked to the underlying population reproduction number, which itself is robust to data thinning. This higher reproduction number is critical for informing control measures as it identifies the groups with highest spread of infection. While detailed studies of transmission with recording of all cases remains the gold standard for estimating epidemiological parameters, such studies are rarely undertaken during outbreak situations. The methods presented here allow us to use data that are readily collected to test for differences in disease spread between two types of individuals, and give us an estimate of the reproduction number for the high-transmission group that is relatively free of bias.

**APPENDIX A**

We give further details of calculations for the four reproduction matrix forms used in the main text, and outline alternative matrix forms that are less informative. For simplicity of notation, we assume throughout that the contact matrix forms used in the main text, and refer to the corresponding reproduction numbers as R_A and R_C.

For the White & Pagano method [9,10], define the daily case numbers to be A_t, and set C_t = \sum_{t=1}^{T} A_t, \quad \hat{C} = \sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} C_{t-j} w_j, \quad \hat{A} = \sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} A_{t-j} w_j.

In order to calculate unambiguous estimates of the reproduction numbers, we assume

\[ \frac{\sum_{j=1}^{\min(k,t)} A_{t-j} w_j}{\sum_{j=1}^{\min(k,t)} C_{t-j} w_j} \approx \frac{\sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} A_{t-j} w_j}{\sum_{t=1}^{T} \sum_{j=1}^{\min(k,t)} C_{t-j} w_j} = \frac{\hat{A}}{\hat{C}}, \]

that is, that the proportion of infected individuals in each class is approximately constant over time. Under these conditions, the following equations hold

\[ \hat{C} = m_{CC} \hat{C} + m_{CA} \hat{A}, \]

and

\[ \hat{A} = m_{AC} \hat{C} + m_{AA} \hat{A}. \]

If the proportion of individuals in each class does change over the course of the outbreak, then it is likely that the reproduction numbers also change over time, and the Wallinga & Teunis method [7] should be used instead.

For the Wallinga & Teunis method [7], we define \( f = C/(A + C) \) to be the fraction of cases that are children, either for the entire outbreak or daily as \( f(t) \). When assessing data thinning, we assume that the population reproduction number, \( R \), is relatively insensitive to thinning [26], and assess the sensitivity of the two reproduction numbers to errors in the estimate of \( f \).

In the case of the symmetric matrices (separable and HiC2C), we assume \( f \geq 0.5 \), in the case of the proportional matrix, we assume \( f \geq p \) and in the case of the contact-frequency matrix, we assume \( (1 - q_A)/(1 - q_A) > 1 \) to ensure that \( R_C \geq R_A \).

**A.1. Separable matrix**

The reproduction matrix has the form \( \begin{bmatrix} a^2 & ab \\ ab & b^2 \end{bmatrix} \). White & Pagano. We have

\[ a = \frac{\hat{C}}{\sqrt{\hat{C} + \hat{A}}}, \quad b = \frac{\hat{A}}{\sqrt{\hat{C} + \hat{A}}} \]

and

\[ R_C = \frac{\hat{C}(\hat{C} + \hat{A})}{\hat{C} + \hat{A}}, \quad R_A = \frac{\hat{A}(\hat{C} + \hat{A})}{\hat{C} + \hat{A}}. \]

Wallinga & Teunis. The dominant eigenvalue of this matrix is \( R = a^2 + b^2 \) and we set \( x = (1 - f)/f \). By definition, \( q_C = a^2 f/(a^2 f + ab(1 - f)) = a^2/f \) and similarly, \( q_A = b^2/f \). Further, \( a^2 f + ab(1 - f) = fR \) and \( ab f + b^2 (1 - f) = (1 - f)R \), which simplifies to \( a^2 + abx = R \) and \( ab + b^2 x = xR \). Some algebra then gives us \( q_C = 1/(1 + x^2), \quad q_A = x^2/(1 + x^2), \quad R_C = (R(1 + x))/(1 + x^2) \) and \( R_A = (R(1 + x))/(1 + x^2) \) and the reproduction matrix has form

\[ R = \frac{R}{(1 + x^2)} \begin{bmatrix} 1 & x \\ x & x \end{bmatrix}. \]

**Sensitivity to thinning.** \( R_C \) is maximized at \( f = 1/\sqrt{2} \approx 0.71 \), and we assume \( f \geq 0.5 \), which implies that \( R_C \in [R,1.21] \), since \( R_C(f=1/2) = R_C(f=1) = R \) and \( R_C(f=1/\sqrt{2}) \approx 1.21R \). In contrast, \( R_A \in [R,0.1] \).

**A.2. HiC2C matrix**

The reproduction matrix has the form \( \begin{bmatrix} a & b \\ b & c \end{bmatrix} \). White & Pagano. We have

\[ a = \frac{\hat{C}(\hat{C} + \hat{A}) - \hat{A}^2}{\hat{C}(\hat{C} + \hat{A})} \quad \text{and} \quad b = \frac{\hat{A}}{\hat{C}(\hat{C} + \hat{A})}, \]

and the reproduction numbers are

\[ R_C = \frac{\hat{C}(\hat{C} + \hat{A}) + \hat{A}(\hat{C} - \hat{A})}{\hat{C}(\hat{C} + \hat{A})} \quad \text{and} \quad R_A = \frac{2\hat{A}}{\hat{C} + \hat{A}}. \]

Wallinga and Teunis. We have \( q_C = 1 - (1 - f)^2/f \) and \( q_A = 1 - f \). The mean reproduction numbers are:

\[ R_C = \frac{R(-2f^2 + 4f - 1)}{f} \]

\[ R_A = 2R(1 - f) \]

and reproduction matrix: \( R \begin{bmatrix} q_C & q_A \\ q_A & q_A \end{bmatrix} \).
Sensitivity to thinning. As above, \( R_C \) is maximized at \( f = 1/\sqrt{2} \approx 0.71 \). We assume \( f \geq 0.5 \), which implies that \( R_C \in R[1,1.7] \), since \( R_C(f = 1/2) = R_C(f = 1) = R \) and \( R_C(1/\sqrt{2}) = 2R(2-\sqrt{2}) \approx 1.17R \). Again, \( R_A \in R[0,1] \).

A.3. Contact frequency

The reproduction matrix has the form

\[
\begin{bmatrix}
\rho_C a^2 & (1 - \rho_A)ab \\
(1 - \rho_C)ab & \rho_A b^2
\end{bmatrix},
\]

where \( \rho_C \) is the fraction of a child’s contacts that are with children, and \( \rho_A \) is the fraction of an adult’s contacts that are with adults.

White & Pagano. We have

\[
b = \frac{\hat{C} - \rho_C a^2 \hat{C}}{(1 - \rho_A) a}\,
\]

where \( a \) is the solution of

\[
\rho_C \hat{C}^2 (\rho_A + \rho_C - 1) a^2 + [(1 - \rho_C - \rho_A - \rho_A \rho_C) \hat{C}^2 - (1 - \rho_A)^2 \hat{A} a^2 + \hat{C}^2 \rho_A = 0.
\]

Wallinga & Teunis. We have

\[
q_C = 1 - (1 - q_C) \frac{f^2(1 - \rho_C)}{(1 - f)^2(1 - \rho_A)},
\]

where \( q_C \) is the solution between 0 and 1 of

\[
(\rho_A + \rho_C - 1) q_C^2 + (1 - \rho_C - \rho_A - \rho_A \rho_C) q_C + \rho_A \rho_C = 0.
\]

Sensitivity to thinning. Analysing the effect of thinning on the reproduction numbers is more complicated for the contact-frequency case as there are two parameters in the model. However, if we fix them at \( \rho_A = 0.75 \) and \( \rho_C = 0.5 \) (as throughout this paper), we find that \( R_C \in R[1,1.29] \) and \( R_A \in R[0,1] \).

A.4. Proportional

The reproduction matrix has the form

\[
\begin{bmatrix}
a + bp & bp \\
bp & b(1 - p)
\end{bmatrix},
\]

where \( p \) is the fraction of children in the population, assumed to be known.

White & Pagano. We have

\[
a = \frac{\hat{C} - \rho_C a^2 \hat{A}}{\hat{C}}, \quad b = \frac{\hat{A}}{(1 - p)(\hat{C} + A)}
\]

and

\[
R_C = \frac{(1 - p)(\hat{C} + \hat{A} + \hat{A} \hat{C} - p \hat{A} \hat{A})}{\hat{C}(1 - p)(\hat{C} + A)}
\]

and

\[
R_A = \frac{\hat{A}}{(1 - p)(\hat{C} + A)}.
\]

Wallinga & Teunis. We have \( q_C = (f - p + fp - f^2 p)/(f(1 - p), \) \( q_A = 1 - f \). The reproduction numbers are

\[
R_C = (R(2f - p - f^2)) / (1 - f), \quad R_A = (R(1 - f)) / (1 - p)\]

and the matrix has the form

\[
R = \begin{bmatrix}
(1 - f)p & (1 - f)p \\
(1 - p)(1 - f) & (1 - p)(1 - f)
\end{bmatrix}.
\]

Sensitivity to thinning. Assuming \( f \geq p \), \( R_C \in R[1, K] \), where \( K = (2(1 - \sqrt{p}))/ (1 - p) \). For \( p = 0.25 \), \( K \approx 1.333 \). We also have \( R_C(f = p) = R_C(f = 1) = R \). Once again, \( R_A \in R[0,1] \).

A.5. Type-specific transmission

If the matrix has the form

\[
\begin{bmatrix}
a & b \\
b & a
\end{bmatrix},
\]

then all individuals have the same reproduction number, which is not informative. A matrix of the form

\[
\begin{bmatrix}
a^2 & b \\
b & a
\end{bmatrix}
\]

also places considerable restrictions on the pairs of reproduction numbers that can be estimated—for instance, this form cannot distinguish between the reproduction numbers for adults and children in the Japanese dataset shown in figure 3.

A.6. Lower diagonal matrix

An alternative asymmetric matrix can be constructed by assuming that there is no transmission from adults to children, giving a matrix of the form

\[
\begin{bmatrix}
a & 0 \\
b & b
\end{bmatrix}.
\]

This format gives results similar to the two asymmetric matrices outlined above, but is less flexible.

A.7. Varying infectivity or susceptibility

If infectivity varies with type, the matrix has the form

\[
\begin{bmatrix}
a & b \\
a & b
\end{bmatrix}.
\]

Then, we have \( f = 0.5 \) for all values of \( a \) and \( b \). Although the reproduction numbers are different (children have
reproduction number \( R_0 \), and adults, \( R_1 \), we cannot observe these differences from the data.

If susceptibility varies with type, the matrix has the form

\[
\begin{pmatrix}
a & a \\ b & b
\end{pmatrix}
\]

and all individuals have reproduction number \( (a + b) \).

REFERENCES


