



The correlation between infectivity and incubation period of measles, estimated from households with two cases

Don Klinkenberg, Hiroshi Nishiura

► To cite this version:

Don Klinkenberg, Hiroshi Nishiura. The correlation between infectivity and incubation period of measles, estimated from households with two cases. *Journal of Theoretical Biology*, 2011, 284 (1), pp.52. 10.1016/j.jtbi.2011.06.015 . hal-00720863

HAL Id: hal-00720863

<https://hal.science/hal-00720863>

Submitted on 26 Jul 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Author's Accepted Manuscript

The correlation between infectivity and incubation period of measles, estimated from households with two cases

Don Klinkenberg, Hiroshi Nishiura

PII: S0022-5193(11)00314-6
DOI: doi:10.1016/j.jtbi.2011.06.015
Reference: YJTBI6517



www.elsevier.com/locate/jtbi

To appear in: *Journal of Theoretical Biology*

Received date: 21 December 2010
Revised date: 14 June 2011
Accepted date: 15 June 2011

Cite this article as: Don Klinkenberg and Hiroshi Nishiura, The correlation between infectivity and incubation period of measles, estimated from households with two cases, *Journal of Theoretical Biology*, doi:[10.1016/j.jtbi.2011.06.015](https://doi.org/10.1016/j.jtbi.2011.06.015)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1 The correlation between infectivity and incubation period of
2 measles, estimated from households with two cases

3 Don Klinkenberg^{a*}, Hiroshi Nishiura^{a,b,1}

4 ^a. *Theoretical Epidemiology, Department of Farm Animal Health, Faculty of Veterinary*
5 *Medicine, Utrecht University, Yalelaan 7, 3584 CL Utrecht, The Netherlands;*

6 ^b. *PRESTO, Japan Science and Technology Agency, Saitama, Japan*

7 ¹. *Present address: School of Public Health, The University of Hong Kong, Pokfulam, Hong*
8 *Kong Special Administrative Region*

9 * d.klinkenberg@uu.nl

10

11 ABSTRACT

12 The generation time of an infectious disease is the time between infection of a primary case
 13 and infection of a secondary case by the primary case. Its distribution plays a key role in
 14 understanding the dynamics of infectious diseases in populations, e.g. in estimating the basic
 15 reproduction number. Moreover, the generation time and incubation period distributions
 16 together characterize the effectiveness of control by isolation and quarantine. In modelling
 17 studies, a relation between the two is often not made specific, but a correlation is biologically
 18 plausible. However, it is difficult to establish such correlation, because of the unobservable
 19 nature of infection events. We have quantified a joint distribution of generation time and
 20 incubation period by a novel estimation method for household data with two susceptible
 21 individuals, consisting of time intervals between disease onsets of two measles cases. We
 22 used two such datasets, and a separate incubation period dataset. Results indicate that the
 23 mean incubation period and the generation time of measles are positively correlated, and that
 24 both lie in the range of 11-12 days, suggesting that infectiousness of measles cases increases
 25 significantly around the time of symptom onset. The correlation between times from infection
 26 to secondary transmission and to symptom onset could critically affect the predicted
 27 effectiveness of isolation and quarantine.

29 HIGHLIGHTS

- 30 • Household measles data, consisting of time intervals between symptom onset of two
 31 cases, are analysed
- 32 • Correlation between the times from infection to symptom onset (incubation time) and
 33 to secondary transmission (generation time) are explicitly modelled
- 34 • Measles incubation time and generation time appear positively correlated, which
 35 could affect the effectiveness of case isolation and quarantine

36 **Keywords:** serial interval; incubation period; transmission; epidemiology; measles

37

Accepted manuscript

38 1. Introduction

39 To capture the transmission dynamics of infectious diseases appropriately, a good
 40 understanding of specific time events and intervals concerning infection processes and onset
 41 phenomena is crucial [1-4]. Of the various time intervals describing the intrinsic transmission
 42 process, the generation time is defined as the time between infection of a primary case and
 43 infection of secondary cases caused by the primary case [5]. Mathematical theory describing
 44 the increase in infected individuals shares its origin with that of mathematical demography
 45 [6], where the distribution of the generation time (which describes the time from birth to the
 46 age at reproduction) has been known to play a key role in describing population dynamics by
 47 successive generations of birth events. Similarly, for infectious diseases, it is essential to
 48 know the generation time distribution to offer a robust estimate of the reproduction number (a
 49 measure for the transmission potential of a disease) from real-time epidemic growth data [3, 7,
 50 8].

51 A second important interval is the incubation period, defined as the time between
 52 infection and onset of symptoms. Knowing the incubation period distribution is crucial for
 53 control of an infectious disease, e.g. to inform how long people should be quarantined, or
 54 how long contacts should be traced back to find other cases. The generation time and
 55 incubation period distributions together determine the effectiveness of control measures such
 56 as ‘transmission-reducing’ treatment, isolation and quarantine [9, 10], through the level of
 57 infectivity before onset of symptoms. Current calculations, however, do not make specific
 58 assumptions on dependence between generation time and incubation period. Dependency
 59 may be expected from a biological perspective and could affect the effectiveness of isolation.
 60 More quantitative knowledge on the relationship between infectivity and disease is therefore
 61 essential.

62 Whereas for the incubation period, data are often available from persons with a single

63 and known contact to an infected, it is not an easy task to estimate the generation time
64 distribution in practice, because infection events are seldom directly observable. There have
65 been attempts, for example, to estimate the generation time distribution from viral shedding
66 data [11], but it is difficult to link virological findings to the epidemiological phenomenon of
67 secondary transmission without further information (e.g. frequency, mode and degree of
68 contacts).

69 Previously, the distribution of the generation time was implicitly assumed to correspond
70 exactly to that of the serial interval [12], defined as the time from symptom onset in a given
71 case to symptom onset in secondary cases [13-15]. The relation between the two intervals is
72 shown in Figure 1. It shows that the serial interval between two cases δ is in fact an
73 aggregation of three intervals, namely the generation time u and two incubation periods t_1 and
74 t_2 . Here we will assume that t_1 and t_2 are equally distributed, although even this is not obvious
75 if infectiousness and symptoms are correlated. Under this assumption, it is clear that the mean
76 of the generation time and serial interval distributions should be the same, but implicitly
77 assuming that the generation time distribution corresponds exactly to the serial interval
78 distribution is incorrect and may lead to flawed estimation of the reproduction number. A
79 correction of the variance by use of incubation period data could be considered, but is not
80 straightforward, because of the likely dependence between the generation time and incubation
81 period of the primary case.

82 Estimation of the generation time distribution by use of observed intervals between
83 disease onset of two cases, is complicated for more reasons. For instance, it may be that
84 asymptomatic cases are missed so that two cases are not uniquely or only indirectly linked
85 [16]. Also, the secondary case may develop symptoms before the primary case, resulting in a
86 negative serial interval but not observable as such. Finally, cases may have been infected both
87 by an external source (co-primary cases), either at the same time or at different and

independent times. We will in the rest of the paper use d to denote any interval between disease onset of two cases, and reserve δ for the theoretical (real) serial interval. When analyzing such data, all possible mechanisms should be accounted for.

To tackle this issue, the present study revisits the time interval between onsets of two measles cases in households (i.e. distribution of the time interval between onset of the first case and onset of the second case in families with two susceptible individuals). This type of data, with the characteristic bimodal distribution, has been studied and reviewed extensively [14, 17-19]. In the present study, we propose a new method to estimate the generation time distribution from these data, and we aim at explicitly quantifying the correlation between the generation time and incubation period. Since household contacts are made in intense conditions of close proximity [15], we realize that our estimates cannot directly be extrapolated beyond the household setting. Rather, we aim – for the first time – at specifically addressing the dependence between generation time and incubation period, and at developing a methodological basis to estimate the generation time distribution using observable epidemiological data. In addition, we address identifiability of the model parameters, in particular the correlation coefficient, by analysis of datasets simulated with parameter estimates.

2. Materials and methods

2.1. Data

We analyzed time intervals between onset of first and second cases of measles in households with two susceptible individuals. Here we use the terms “first” and “second” cases (i.e. not primary and secondary cases), because it is unknown whether the second case was infected by the first, the first by the second, or whether they were both infected in the community. While many analyses of the serial interval of measles were performed with Hope

Simpson's [20] or Kenyan datasets [21], the present study used two datasets with larger sample sizes which were collected in Rhode Island from 1917-23 [22] and 1929-34 [23] and were previously revisited to estimate latent and infectious periods [24]. Figure 2 shows the observed distributions extracted from the two abovementioned publications, the original data of which are available in Appendix A of the Supplementary Materials.

In total, 5,762 and 4,516 intervals between two cases in the households were recorded in 1917-23 and 1929-34, respectively. In the former dataset, there were no documented pairs of cases with $d = 0$ (where d is the observed interval in days between onset of the two cases), but there are many pairs with $d = 1$. Following Wilson *and others* [23], we assume that all pairs with $d = 0$ were included in this dataset as $d = 1$. In addition, although both datasets included pairs with $d > 30$ days, all observations of $d > 30$ were grouped together.

In addition to the time intervals between two onsets of measles in a household, we extracted a separate incubation period dataset from a different historical publication of measles in 1931 [25, 26] for use in our analyses. In that study, incubation periods were recorded for 116 cases based on detailed observations of contact which arose from cases that experienced only a single day of exposure. Figure 3 illustrates the incubation period distribution, the original data of which are available in Appendix B of the Supplementary Materials. Sample mean (median) and variance of the incubation period were 12.3 (12.0) days and 12.1 days² respectively.

2.2. Theoretical basis

Here we explain the epidemiological mechanisms underlying the bimodal distribution of the time intervals in Figure 2. As noted previously [27, 28], two cases in a household with two susceptible individuals are the result of either (i) community infection in both cases (i.e. "co-primary" cases) or (ii) community infection of the first case and household infection of

the second. Possibility (i) is then reflected in the first peak and possibility (ii) corresponds the second peak. Although previous reviews have discussed this issue implicitly by defining a cut-off point of d to distinguish (i) from (ii) [14, 18], we do not take this approach and, rather, employ a mixture distribution. In addition, we allow for a third possible infection history, namely that both cases were independently infected in the community. Figure 1(a) illustrates the three mechanisms leading to $f(d)$, the probability distribution of d :

Mechanism 1. Both cases independently experienced infection in the community, which accounts for a proportion π_1 of the observations (where $0 \leq \pi_1 \leq 1$).

Mechanism 2. Both cases were infected at an identical point in time in the community, which accounts for a proportion π_2 (where $0 \leq \pi_2 \leq 1 - \pi_1$).

Mechanism 3. One case was infected in the community and the other was infected by the primary case in the household, which accounts for a proportion $1 - \pi_1 - \pi_2$.

If we denote the probability of difference d by mechanisms 1, 2 and 3 by $q_1(d)$, $q_2(d)$ and $q_3(d)$, respectively, we can express the mixture density $f(d)$ as

$$f(d) = \pi_1 (q_1(d) + q_1(-d)) + \pi_2 (q_2(d) + q_2(-d)) + (1 - \pi_1 - \pi_2) (q_3(d) + q_3(-d)) \quad (2.1)$$

That is, we assume that the bimodal distribution in Figure 1 is decomposed into three underlying epidemiological mechanisms, expressed as a three-component mixture distribution. The densities q_1 and q_2 are assumed to reflect the difference in onset between the two cases in random order, which is why all densities q_1 , q_2 , and q_3 , d can be negative and positive. In $f(d)$, the difference $d \geq 0$.

2.3. The model

Here we describe how the three mechanisms result in three probability distributions, specifying our parameters of interest. The first term of the mixture distribution, $q_1(d)$, in

equation (2.1) is

$$q_1(d) = p_{(0, \sigma_1^2)}(d), \quad (2.2)$$

which is the density of a normal distribution with mean 0 and variance σ_1^2 . This distribution arises if measles occur in (seasonal) epidemics, and the dataset does not contain pairs of cases occurring in different epidemics. Then, the time interval d is the time difference between onset of two independently sampled cases from an epidemic. Assuming that each epidemic curve is approximately bell-shaped (Farr's model [29]), d is normally distributed with twice the variance of the epidemic curve.

The probability of observing a difference d owing to mechanism 2 depends on the density function $g(t)$ of the incubation period of length t . Because the two cases are infected at the same time, d is the difference (i.e. cross-correlation) of independently and identically distributed incubation periods. Thus, the probability density function $q_2(d)$ is equal to

$$q_2(d) = \int_0^\infty g(\tau) g(d + \tau) d\tau \quad (2.3)$$

Because either case can be the first to show symptoms, a negative d should be regarded as one case being the first (e.g. the oldest), and a positive d as the other being the first.

Mechanism 3 reflects the serial interval within households. From Figure 1 we see that the serial interval δ is given by

$$\delta = u + t_2 - t_1, \quad (2.4)$$

which is interpreted as the sum of the generation time u and incubation period of the secondary case t_2 minus the incubation period of the primary case t_1 . Thus, if u , t_1 and t_2 were independent random variables, the serial interval distribution would be the convolution of the generation time and incubation period distributions followed by the cross-correlation of this convolution and the incubation period distribution. However, it is biologically natural to assume that t_1 and u are dependent (e.g. cases with short incubation period may cause

secondary transmission earlier on average than cases with long incubation periods). Therefore, we assume that u and t_1 follow a joint distribution $h(t_1, u)$, so that the probability of observing difference d due to mechanism 3 depends on the incubation period distribution $g(t_2)$ and on this joint distribution $h(t_1, u)$. Under this assumption, $g(t)$ is a marginal distribution of $h(t, u)$. Now, the probability density $q_3(d)$ is equal to

$$q_3(d) = \int_0^\infty \int_0^\infty h(\tau, u) g(d - u + \tau) du d\tau \quad (2.5)$$

2.4. Statistical analysis

We fitted a lognormal, gamma, and Weibull distribution to the incubation period data, by means of maximum likelihood estimation. Akaike's Information Criterion (AIC) was used to decide which of these distributions to choose for further analysis. The Weibull distribution was rejected for further use.

For construction of joint (cumulative) distribution functions $H(t, u)$, we combined two lognormal or gamma marginal distributions by use of a Gaussian copula function. Copula functions provide a correlation structure to sets of marginal distributions [30]. In our case, we first denote the marginal distribution functions of T and U by $P_T(t)$ and $P_U(u)$, respectively. Because $0 < P_T(t), P_U(u) < 1$, we can write the joint distribution as

$$H(t, u) = \Phi_\rho \left(\Phi^{-1}(P_T(t)), \Phi^{-1}(P_U(u)) \right) \quad (2.6)$$

Here, Φ is a normal distribution with standard normal marginals [30]. For the marginal distributions $P_T(t)$ and $P_U(u)$, we chose two lognormal distributions or two gamma distributions (not Weibull, as argued above), as these are the most commonly used for generation times and incubation times [2, 12, 31]. We will refer to these distributions as the

bilognormal and bigamma distributions.

Both distributional assumptions result in distributions with 5 parameters; i.e. two parameters for the marginal incubation period and generation time distributions in addition to a correlation parameter ρ . Since we additionally estimate σ_1 , π_1 and π_2 , there are in total 8 parameters to be estimated.

Because the density of d with the third mechanism (within household transmission) includes a convolution of $g(t)$ and the bivariate density $h(t,u)$, a likelihood approach was computationally too challenging, even in a classic Bayesian procedure where the likelihood need not be maximized but evaluated at many sampled points in parameter space. Therefore, we employed Sequential Monte Carlo sampling (SMC), in particular the ABC-PRC algorithm (Approximate Bayesian Computation – Partial Rejection Control), described by Sisson *and others* [32]. The ABC-PCR algorithm was programmed in R statistical software. A step-by-step explanation of the algorithm as we used it is given in Appendix C of the Supplementary Materials. The idea is as follows: the algorithm starts by sampling 2000 parameter sets from prior distributions. Subsequently, a second population of parameter sets is obtained by sampling from the first, perturbing, simulating a dataset and comparing the simulated data to the real data. If both datasets are close enough, ie if the χ^2 statistic is below some selection threshold ε_2 , the parameter set is accepted for the second population. Then a third population is obtained from the second with a more restrictive selection threshold ε_3 , then a fourth, etc, until the final selection threshold ε_{\min} is met. From the final population of parameter sets, the medians and 95% credible intervals are derived.

The fits of the two distributional assumptions, bilognormal and bigamma, were compared by use of the deviance $-2\sum \left\{ f_i \log(f_i/n_i) - f_i \log(\hat{f}_i/n_i) \right\}$, in which f_i are the observed frequencies, \hat{f}_i the expected frequencies, and n_i the total numbers of observations

which f_i is part from (5762 or 4516 onset differences, or 116 incubation periods). For the expected frequencies of onset differences a simulated distribution was constructed with 100,000 samples with the median parameter estimates. The expected frequencies of the incubation periods were calculated exactly.

2.5. Parameter identifiability

To address parameter identifiability of the proposed estimation method while saving substantial time for the iteration procedure, we have made four different assumptions regarding the data generating process of the ‘real model’. That is, we simulated datasets that were similar to the original data under the four assumptions. The simulated datasets were subsequently analyzed using the two models (bivariate lognormal and bivariate gamma), and the resulting estimates were compared to the parameter values used for the simulations. The four assumptions on the real model were:

- (1) a bivariate lognormal distribution of incubation and generation times with positive correlation (20 simulated datasets)
- (2) a bivariate gamma distribution of incubation and generation times with positive correlation (20 simulated datasets)
- (3) a bivariate lognormal distribution of incubation and generation times without correlation (10 simulated datasets)
- (4) a bivariate gamma distribution of incubation and generation times without correlation (10 simulated datasets)

Simulations with (1) and (2) were carried out with consensus parameter values based on the four sets of original estimates derived from the two empirical datasets. For simulations with (3) and (4), consensus parameters were obtained after re-analysis of the original datasets, forcing the correlation coefficient at 0 (no correlation), with both lognormal and gamma

models.

Identifiability was first assessed for each parameter separately. The mean of the posterior medians was calculated, and we determined how frequent the median was lower and higher than the true value. In addition, we determined how frequent the 95% credible interval contained the true parameter value, and we tested whether this was significantly different from 95% (Fisher's exact test). Identifiability for the whole model was quantified by the number of parameters failing the latter test (number of parameters correctly identified).

3. Results

3.1. Incubation period data

The fitted lognormal, gamma, and Weibull distributions are shown in Figure 3. The AICs for the three models were 605.7, 608.2, and 626.1, respectively. Thus, we decided not to continue with the Weibull distribution.

3.2. Bilognormal distribution

Assuming a bilognormal distribution for $h(t,u)$, the SMC algorithm took 22 and 21 iterations before reaching ϵ_{\min} for the two datasets, respectively. The resulting parameters and 95% percentile intervals are presented in Table 1, where the actual estimates of the distributional parameters have been translated into means and standard deviations of the intervals of Figure 1. The mean incubation periods of 11.6 and 11.3 days were very similar to the mean generation times of 12.2 and 11.2 days in the years 1917-1923 and 1929-1934, respectively. This is also reflected in the mean time from onset to secondary transmission, which is close to 0. As expected, the mean serial intervals were equal to the mean generation times, but the standard deviations of the serial intervals are smaller. Both datasets indicate a strong and

significant correlation between the generation time and the incubation period, with estimated correlation coefficients of 0.903 and 0.969. Figure 4 shows the densities of the bivariate distributions of incubation and generation times according to the point estimates. Figure 2 shows the theoretical distributions of the difference between the days of onset, indicating the contributions of the three mechanisms. It appears that the more irregular 1917-1923 dataset (deviance = 259) results in a slightly worse fit than the 1929-1934 data (deviance = 157).

3.3. Bigamma distribution

With the bigamma distribution for $h(t,u)$, the SMC algorithm took 12 and 13 iterations before reaching ε_{\min} for the two datasets, respectively. The results are given in Table 2. Compared to the bilognormal model, the standard deviations of the incubation period and generation time were smaller. Unlike the bilognormal model, the serial intervals now have larger standard deviations than the generation times. The correlation coefficient ρ was smaller as well, and with the first dataset, ρ was not even significantly larger than 0, but with the second dataset it was. With the bigamma distribution, the 1929-1934 data (deviance = 188) fitted better than the 1917-1923 data (deviance = 247), as was the case with the bilognormal distribution.

Comparing the two distributional assumptions, the bigamma distribution fitted better with the 1917-1923 data, but the bilognormal distribution with the 1929-1934 data. Figure 4 shows estimated bilognormal and bigamma distributions. The contributions of the three mechanisms was also different between the two distributions, but Figure 2 indicates that in all cases the first peak, second peak, and fat tail of the onset difference distribution are clearly explained by the three mechanisms separately.

3.4. Parameter identifiability

Table 3 shows a summary of the identifiability analysis, the details of which, including

parameter values used for simulation, are given in Appendix D of the Supplementary Material. It turns out that both lognormal and gamma models correctly identify the three components of the mixture distribution, with the bigamma method giving reliable estimates of the proportions of the components in most cases. Regarding the overall performance of the two models, the bivariate gamma model produces more reliable results: with three groups of simulated datasets, all 95% CIs for the eight parameters contained the true values (Table 3). The bivariate lognormal model correctly identified maximally 6/8, in only one case.

The correlation coefficient is significantly overestimated by the lognormal model, but correctly identified by the gamma model, although slightly underestimated. The standard deviation of the generation time is also significantly overestimated by the lognormal model, and slightly underestimated by the gamma model. All other parameters are correctly identified or only slightly biased. In summary: only the gamma model produces reliable estimates, especially on the correlation between the incubation and generation times.

4. Discussion

The present study revisited the distribution of the time interval between the onsets of two cases of measles in households with two susceptible individuals. We decomposed the observed bimodal distribution into a mixture of three distributions, each explicitly linked to a relevant underlying epidemiological mechanism. We focused in particular on the third mechanism which reflects the time interval required for household transmission between first and second cases. Clearly illustrating the relationship between the incubation period, generation time and serial interval (Figure 1), we developed a method to estimate the generation time distribution, addressing dependence between the generation time and incubation period. To the best of our knowledge, the present study is the first to clearly decompose the observed data into underlying epidemiological mechanisms with realistic

definitions of the intrinsic parameters (without arbitrary cut-off values), reasonably deriving parameter estimates for the generation time distribution.

We have used two datasets for two periods of observation, and a third dataset with incubation period observations. This third dataset is required to obtain estimates for the marginal incubation period distribution, especially for the mean because the onset difference datasets only contain differences between two incubation periods, and those have a mean of 0 by definition. Because the first peak, reflecting the mechanism of dependent community transmission, should contain information on the variance of the incubation period, we tried to carry out the analysis without the incubation period data. That was unsuccessful, because the three mechanisms did explain the onset difference distribution incorrectly: the first peak was explained by mechanism 1 (independent infection outside household) with small σ_1 , whereas the incubation period distribution was estimated to have a short mean and very large variance, thus explaining the fat tail. Fixing the mean incubation period at 12.3 days (the sample mean) so that only the variance needed to be estimated from the onset difference data, did not solve this problem.

Ideally, the separate incubation period dataset would have been sampled from the same population as the household data (or even better: from the same households), because incubation periods could differ between situations. By taking a dataset from the same period and a similar population (UK) we have tried to minimize this issue. The results show that, because of the small number of cases in the dataset (116 cases), the parameter estimates for the incubation period distributions are still sensitive to the household data: in fact, the incubation period data could be considered as prior information for the household analysis.

The contributions of the three mechanisms to the mixture distribution will not be the same in separate datasets. For instance, a large epidemic with more community transmission will result in more households with mechanism 1 (independent infection in the community),

whereas a higher infectiousness and contact rate within the household is related to more observations of mechanism 3 (within-household transmission). The fact that so many factors play a role, makes it difficult to use other type of data in the same epidemic to infer on these mixture probabilities. Fortunately, the three components of the mixture distribution could be well identified by our method, and the proportions of the three mixture components reasonably well estimated, at least by the bigamma method.

In fact, the identifiability analysis clearly showed that the bigamma model in general provided much more reliable results than the bilognormal model. With the original data as well as the simulated data, the bigamma distribution resulted in a smaller correlation coefficient combined with smaller standard deviations of the incubation and generation times. This can be understood when realizing what characteristic of the data is indicative of correlation: the variance of the onset differences resulting from household transmission (mechanism 3), ie the width of the second peak. These onset differences are the sum of one independent incubation time T_2 and the difference of a dependent generation time U and incubation time T_1 . Denoting the single variances by s^2_T and s^2_U , the variance of d_3 is therefore $\text{var}(d_3) = 2s^2_T + s^2_U - 2\rho s_T s_U$. This means that a particular $\text{var}(d_3)$ in the data can be a result of either a high correlation coefficient ρ in combination with high s^2_T (as suggested by the bilognormal model), or of a lower ρ and smaller s^2_T and s^2_U (suggested by the bigamma model). The identifiability analysis has shown that the bilognormal model incorrectly favours the first hypothesis, and that the gamma model can correctly distinguish between the two.

Summarizing the results that were obtained for the joint distribution function of the generation time and incubation period, three findings are notable: (1) the mean incubation period and mean generation time were very close, ranging from 11-12 days; (2) there was a positive correlation between the generation time and incubation period, although this was not

significant with the bigamma distribution for the 1917-23 data; and (3) compared with estimates derived from datasets from 1917-23, estimates given by the dataset from 1929-34 suggest shorter generation times and a stronger correlation between the generation time and incubation period. Findings (1) and (2) reflect transmission under intense contact conditions of close and prolonged physical proximity in the households [15]. Given this close contact and our estimates indicating that most secondary transmission of measles must have occurred before or just after the onset of illness, the infectiousness seems to increase fast just before symptom onset. With similar reasoning, finding (3) could indicate higher transmissibility of measles for the observations from 1929-34 compared to those from 1917-23, either due to more intense contacts or to higher or differently timed virus shedding levels, e.g. because of a different genotype [33].

In recent studies the serial interval δ was decomposed as the sum of time from the onset of a primary case to secondary transmission, w , and incubation period of secondary cases, t_2 [34-36]. That is, from Figure 1,

$$\delta = w + t_2 \quad (4.1)$$

has been assumed and further explored. Considering the distributions for each, the relationship is also expressed as

$$f(\delta) = \int_{-x}^{\delta} w(\delta - \tau) g(\tau) d\tau \quad (4.2)$$

where x is the potentially contagious period before onset of illness in the primary case. The function of public health interest has been focused on $w(\delta - \tau)$, as the distribution can suggest the latest time at which symptomatic cases need to be placed in isolation [37]. Our approach does provide a distribution for w , with $\mu_w = \mu_U - \mu_T$ and σ_w as estimated, but it is more informative as it also provides a distribution of the generation time, related to the important growth rate of an epidemic, and it shows the correlation between disease onset and

infectivity.

Our datasets are limited to cases with two susceptibles in one household and both cases infected. This is not at all a random sample of person-to-person transmission events during a measles outbreak, and does not indicate either how likely transmission is under similar circumstances. However, this does not affect the validity of our two main conclusions that we can draw for measles, first that a positive correlation between incubation period and start of infectiousness is very likely, and second that the infectiousness rises considerably around symptom onset. For control of measles, the correlation between disease and infectivity imply that isolation of cases and quarantine of traced contacts will be even more successful than already predicted [9, 10].

As for future implications, we hope our method will be useful for estimating the generation time of various diseases from well-recorded epidemiological data. Although the generation time estimated from household transmission data is likely to be shorter than that estimated from community transmission data [8, 15], further estimates in the community (or in a specific group of contacts) could be derived from serial intervals based on contact tracing. We believe further clarification of different generation times (e.g. between transmissions in the community and households) would enhance our understanding of the transmissibility of a disease in a heterogeneously mixing population which may be very useful in discussing and appropriately quantifying the transmission dynamics of directly transmitted diseases (e.g. influenza). Another line of future research is concerned with variations of generation time with time, which could also be addressed, analyzing time-varying serial intervals [38]. Given the various future possibilities, we believe our estimation procedure satisfies a need to translate observed datasets into one of the most important key intrinsic parameters of transmission dynamics.

433 **Appendix A. Supplementary data**

434 Supplementary data associated with this article can be found in the online version at doi: .

435

436 **Acknowledgements**

437 We would like to thank two anonymous reviewers for their useful comments. The work of

438 HN was supported by the JST PRESTO program.

439

Accepted manuscript

440 **Table 1.**

441 Parameter estimates with the bilognormal distribution

Parameter	1917-1923		1929-1934	
Mean incubation period	11.6	(10.7 ; 12.4)	11.3	(10.6 ; 12.0)
SD incubation period	2.75	(2.44 ; 3.13)	2.37	(2.14 ; 2.61)
Mean generation time	12.2	(11.9 ; 12.5)	11.2	(11.1 ; 11.4)
SD generation time	3.62	(2.98 ; 4.41)	3.13	(2.55 ; 3.68)
Mean time onset-secondary transmission	0.6	(-0.2 ; 1.5)	0.0	(-0.7 ; 0.6)
SD time onset-secondary transmission	1.69	(1.09 ; 2.29)	1.08	(0.57 ; 1.61)
Mean serial interval	12.2	(11.9 ; 12.5)	11.2	(11.1 ; 11.4)
SD serial interval	3.26	(3.05 ; 3.48)	2.62	(2.44 ; 2.79)
Correlation coefficient [ρ]	0.903	(0.732 ; 0.981)	0.969	(0.871 ; 0.997)
Pr(mechanism 1) [π_1]	0.138	(0.0945 ; 0.205)	0.145	(0.0974 ; 0.213)
Pr(mechanism 2) [π_2]	0.238	(0.200 ; 0.273)	0.189	(0.159 ; 0.220)
SD mechanism 1 [σ_1]	20.3	(17.1 ; 24.4)	16.6	(14.4 ; 19.4)
Deviance	259		157	

442 NOTE: Given are the posterior medians and 95% credible intervals. SD is the standard
 443 deviation. Mechanism 1 refers to two independent infections outside the household;
 444 mechanism 2 refers to two synchronous infections outside the household.

445

446 **Table 2.**

447 Parameter estimates with the bigamma distribution

Parameter	1917-1923		1929-1934	
Mean incubation period	11.5	(10.5 ; 12.5)	11.4	(10.6 ; 12.2)
SD incubation period	2.24	(1.87 ; 2.66)	2.08	(1.69 ; 2.42)
Mean generation time	11.9	(11.5 ; 12.1)	11.1	(10.9 ; 11.3)
SD generation time	1.22	(0.69 ; 2.64)	1.79	(0.79 ; 3.08)
Mean time onset-secondary transmission	0.4	(-0.7 ; 1.5)	-0.3	(-1.1 ; 0.4)
SD time onset-secondary transmission	2.13	(1.41 ; 2.66)	1.29	(0.62 ; 1.91)
Mean serial interval	11.9	(11.5 ; 12.1)	11.1	(10.9 ; 11.3)
SD serial interval	3.08	(2.83 ; 3.36)	2.47	(2.28 ; 2.66)
Correlation coefficient [ρ]	0.413	(-0.511 ; 0.928)	0.837	(0.219 ; 0.982)
Pr(mechanism 1) [π_1]	0.195	(0.132 ; 0.279)	0.188	(0.126 ; 0.253)
Pr(mechanism 2) [π_2]	0.194	(0.153 ; 0.232)	0.162	(0.130 ; 0.195)
SD mechanism 1 [σ_1]	18.0	(15.4 ; 22.1)	15.5	(13.7 ; 18.0)
Deviance	247		188	

448 NOTE: Given are the posterior medians and 95% credible intervals. SD is the standard
449 deviation. Mechanism 1 refers to two independent infections outside the household;
450 mechanism 2 refers to two synchronous infections outside the household.

451

Table 3.

Summary of identifiability analysis.

Estimation model		Simulation model			
		Simulated $\rho = 0.863$		Simulated $\rho = 0$	
		Log-normal	Gamma	Log-normal	Gamma
Lognormal	proportion of correctly identified parameters ^a	6/8	4/8	2/8	3/8
	mean estimated ρ	0.966	0.951	0.801	0.785
	proportion of 95% CI containing real ρ	0/20	1/20	0/10	0/10
Gamma	proportion of correctly identified parameters ^a	6/8	8/8	8/8	8/8
	mean estimated ρ	0.842	0.813	-0.182	-0.253
	proportion of 95% CI containing real ρ	20/20	20/20	10/10	10/10

^a proportion of model parameters of which the true value falls within the 95% CI (sufficiently frequent, Fisher's exact test). Full results in Supplementary Materials.

Figure captions

Fig. 1. Theoretical basis of the observed time intervals between onset of first and second measles case in households with two susceptible individuals. (a) Three underlying mechanisms: 1. both cases are infected independently in the community; 2. both cases are infected at the same time in the community; 3. secondary transmission in the household, ie the serial interval distribution. (b) The relationship between generation time u , incubation times t_1 and t_2 , serial interval δ , and time from onset to transmission w .

Fig. 2. Datasets with onset intervals from 1917-1923 [22] and 1929-1934 [23], together with fitted distributions, separating the contributions of the three mechanisms. Arrows indicate the extreme values where the predicted numbers of cases with mechanism 3 are still above 0. (a) Data from 1917-1923, bilognormal model; (b) Data from 1917-1923, bigamma model; (c) Data from 1929-1934, bilognormal model; (d) Data from 1929-1934, bigamma model.

Fig. 3. Incubation period dataset, extracted from [25], and three fitted distributions.

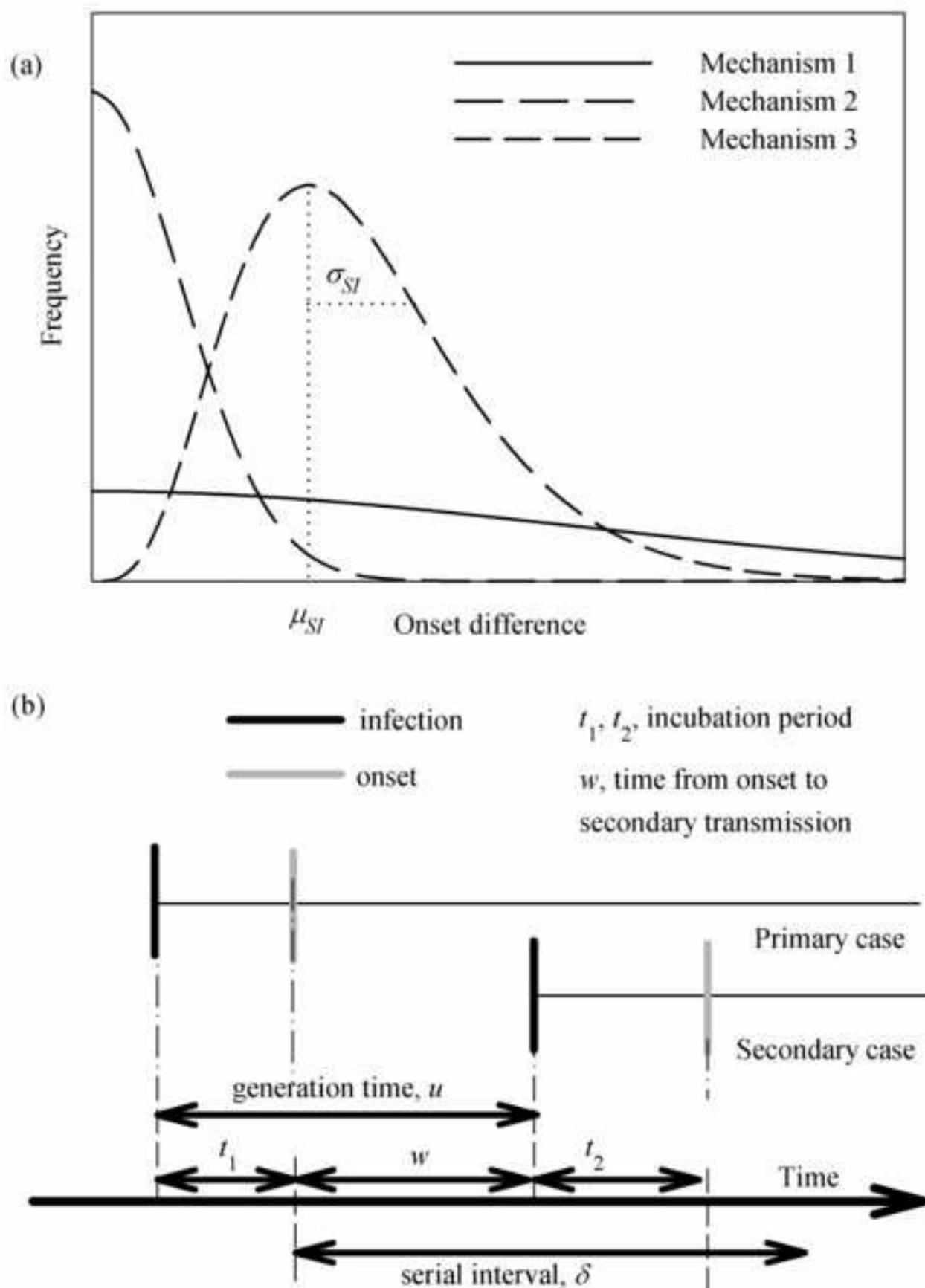
Fig. 4. Joint distribution densities according to median posteriors, for the two datasets and two model. (a)-(d) as in Fig. 2.

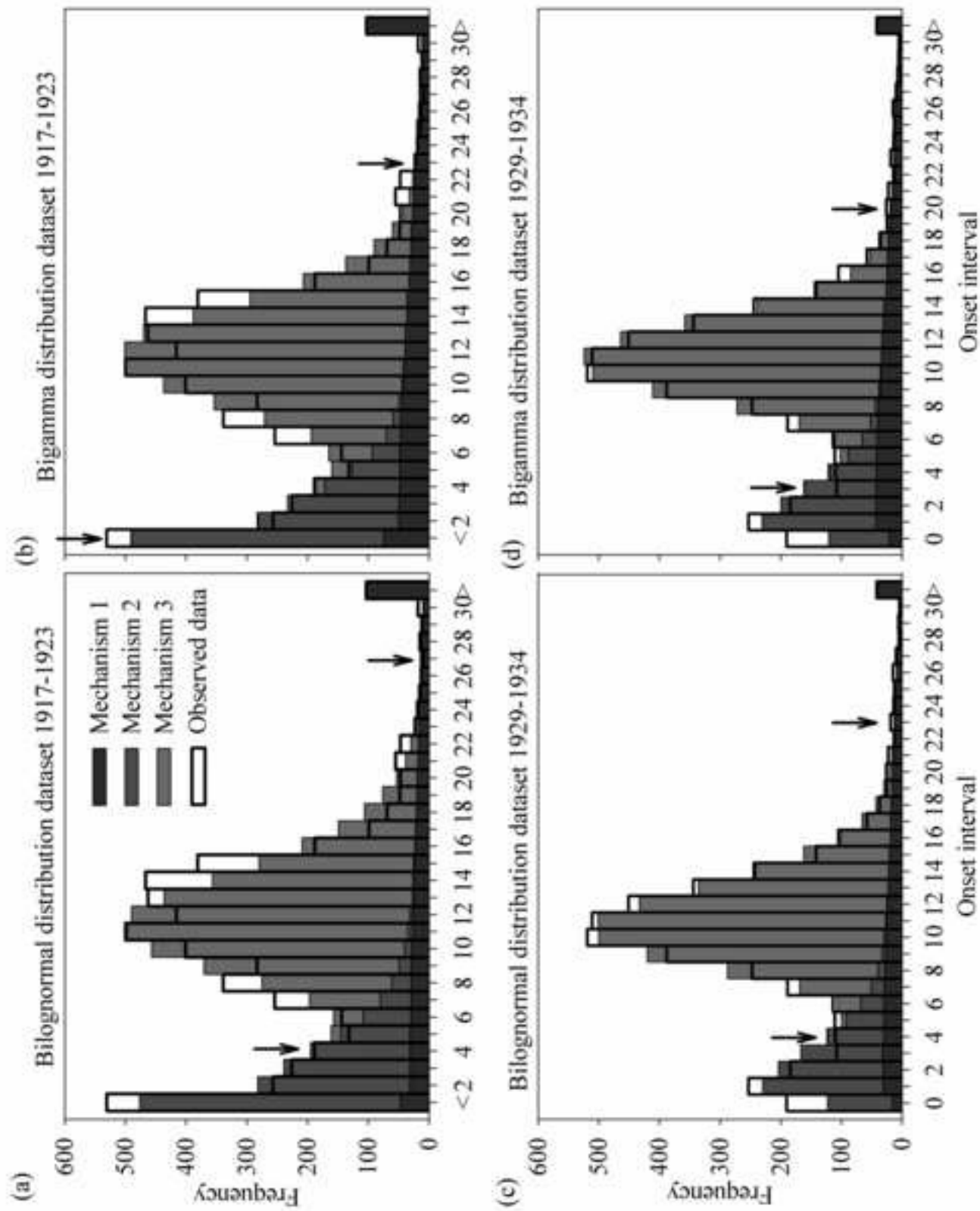
- 477 [1] J.T. Griffin, T. Garske, A.C. Ghani, P.S. Clarke, Joint estimation of the basic reproduction
478 number and generation time parameters for infectious disease outbreaks, *Biostatistics*. in press
479 (2010).
- 480 [2] H.J. Wearing, P. Rohani, M.J. Keeling, Appropriate models for the management of
481 infectious diseases, *PLoS Medicine*. 2 (2005) 621-627.
- 482 [3] J. Wallinga, M. Lipsitch, How generation intervals shape the relationship between growth
483 rates and reproductive numbers, *Proceedings of the Royal Society London B*. 274 (2007)
484 599-604.
- 485 [4] H. Nishiura, Time variations in the transmissibility of pandemic influenza in Prussia,
486 Germany, from 1918-19, *Theoretical Biology and Medical Modelling*. 4 (2007) 20.
- 487 [5] A. Svensson, A note on generation times in epidemic models, *Math Biosci*. 208 (2007)
488 300-311.
- 489 [6] N. Keyfitz, *Applied Mathematical Demography*, John Wiley, New York, USA, 1977.
- 490 [7] M.G. Roberts, J.A.P. Heesterbeek, Model-consistent estimation of the basic reproduction
491 number from the incidence of an emerging infection, *Journal of Mathematical Biology*. 55
492 (2007) 803-816.
- 493 [8] C. Fraser, Estimating individual and household reproduction numbers in an emerging
494 epidemic, *PLoS ONE*. 2 (2007) e758.
- 495 [9] C. Fraser, S. Riley, R.M. Anderson, N.M. Ferguson, Factors that make an infectious disease
496 outbreak controllable, *Proceedings of the National Academy of Sciences*. 101 (2004)
497 6146-6151.
- 498 [10] D. Klinkenberg, C. Fraser, J.A.P. Heesterbeek, The effectiveness of contact tracing in
499 emerging epidemics, *PLoS ONE*. 1 (2006) e12.
- 500 [11] F. Carrat, E. Vergu, N.M. Ferguson, M. Lemaître, S. Cauchemez, S. Leach, A.J. Valleron,
501 Time lines of infection and disease in human influenza: a review of volunteer challenge
502 studies, *American Journal of Epidemiology*. 167 (2008) 775-785.
- 503 [12] J. Wallinga, P. Teunis, Different epidemic curves for Severe Acute Respiratory Syndrome
504 reveal similar impacts of control measures, *American Journal of Epidemiology*. 160 (2004)
505 509-519.
- 506 [13] R.E. Hope Simpson, The period of transmission in certain epidemic diseases. An
507 observational method for its discovery, *Lancet*. 2 (1948) 755-755-760.
- 508 [14] N.T.J. Bailey, *The mathematical theory of infectious diseases and its applications*, 2nd ed.,
509 Charles Griffin & Company Ltd, London, UK, 1975.
- 510 [15] P.E.M. Fine, The interval between successive cases of an infectious disease, *American*
511 *Journal of Epidemiology*. 158 (2003) 1039-1047.

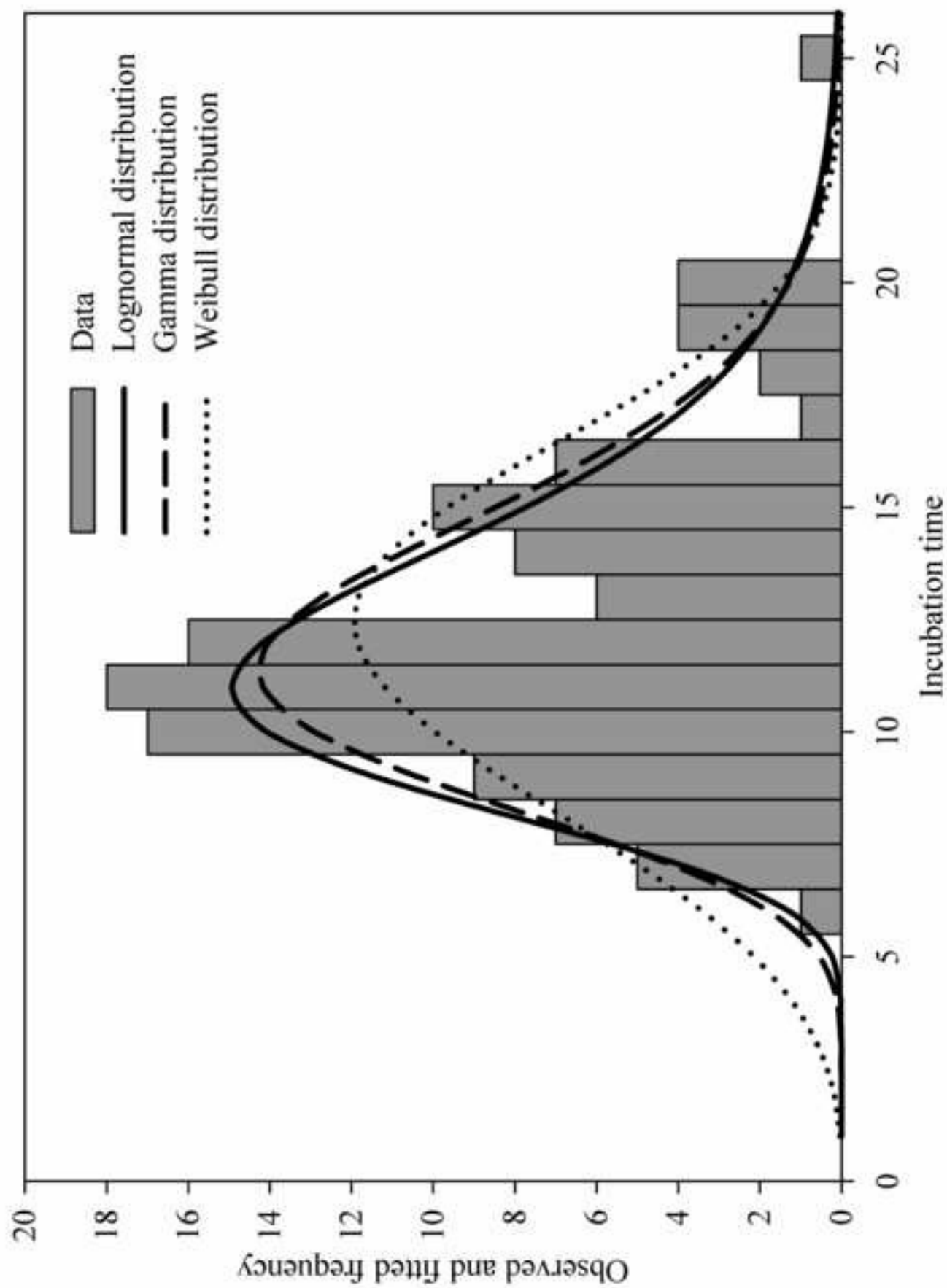
- 512 [16] H. Inaba, H. Nishiura, The state-reproduction number for a multistate class age structured
513 epidemic system and its application to the asymptomatic transmission model, *Math Biosci.* 216
514 (2008) 77-89.
- 515 [17] N.T.J. Bailey, On estimating the latent and infectious periods of measles. I. Families with
516 two susceptibles only, *Biometrika.* 43 (1956) 15-22.
- 517 [18] N.G. Becker, *Analysis of infectious disease data*, Chapman & Hall, Baco Raton, USA,
518 1989.
- 519 [19] P.D. O'Neill, D.J. Balding, N.G. Becker, M. Eerola, D. Mollison, Analyses of infectious
520 disease data from household outbreaks by Markov chain Monte Carlo methods, *Applied*
521 *Statistics.* 49 (2000) 517-542.
- 522 [20] R.E. Hope Simpson, Infectiousness of cimmunicable diseases in the household (measles,
523 chickenpox, and mumps), *Lancet.* 2 (1952) 549-554.
- 524 [21] P. Aaby, J. Leeuwenburg, Patterns of transmission and severity of measles infection: a
525 reanalysis of data from the Machakos Area, Kenya, *J Infect Dis.* 161 (1990) 171-174.
- 526 [22] C.V. Chapin, Measles in Providence, R.II, 1858-1923, *Am J Hyg.* 5 (1925) 635-655.
- 527 [23] E.B. Wilson, C. Bennett, M. Allen, J. Worcester, Measles and scarlet fever in Providence,
528 R.I., 1929-1934 with respect to age and size of family, *Proceedings of the American*
529 *Philosophical Society held at Philadelphia for promoting useful knowledge.* 80 (1939)
530 357-476.
- 531 [24] K.J. Gough, The estimation of latent and infectious periods, *Biometrika.* 64 (1977)
532 559-565.
- 533 [25] E.W. Goodall, Incubation period of measles, *Brit Med J.* 1 (1931) 73-74.
- 534 [26] H. Nishiura, Early efforts in modeling the incubation period of infectious diseases with an
535 acute course of illness, *Emerg Themes Epidemiol.* 4 (2007) 2.
- 536 [27] I.M. Longini, J.S. Koopman, Household and community transmission parameters from
537 final distributions of infections in households, *Biometrics.* 38 (1982) 115-126.
- 538 [28] C.L. Addy, J. Longini I.M., M. Haber, A generalized stochastic model for the analysis of
539 infectious disease final size data, *Biometrics.* 47 (1991) 961-974.
- 540 [29] P.E.M. Fine, John Brownlee and the measurement of infectiousness: an historical study in
541 epidemic theory, *J R Statist Soc A.* 142 (1979) 347-362.
- 542 [30] J. Yan, Enjoy the joy of copulas: with a package copula, *J Stat Soft.* 21 (2007) 1-21.
- 543 [31] B.J. Cowling, M.P. Muller, I.O.L. Wong, L.-. Ho, M. Louie, A. McGeer, G.M. Leung,
544 Alternative methods of estimating an incubation distribution. Examples from severe acute
545 respiratory syndrome, *Epidemiology.* 18 (2007) 253-259.

- 546 [32] S.A. Sisson, Y. Fan, M.M. Tanaka, Sequential Monte Carlo without likelihoods,
547 Proceedings of the National Academy of Sciences. 104 (2007) 1760-1765.
- 548 [33] H.S. El Mubarak, S. Yüksel, G. Van Amerongen, P.G.H. Mulder, M.M. Mukhtar,
549 A.D.M.E. Osterhaus, R.L. De Swart, Infection of cynomolgus macaques (*Macaca fascicularis*)
550 and rhesus macaques (*Macaca mulatta*) with different wild-type measles viruses, J Gen Virol.
551 88 (2007) 2028-2034.
- 552 [34] N.M. Ferguson, D.A.T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S.
553 Iamsrithaworn, D.S. Burke, Strategies for containing an emerging influenza pandemic in
554 southeast Asia, Nature. 437 (2005) 209-214.
- 555 [35] A.Y.C. Kuk, S. Ma, The estimation of SARS incubation distribution from serial interval
556 data using a convolution likelihood, Statistics in Medicine. 24 (2005) 2525-2537.
- 557 [36] H. Nishiura, M. Eichner, Infectiousness of smallpox relative to disease age: estimates
558 based on transmission network and incubation period, Epidemiology and Infection. 135 (2007)
559 1145-1150.
- 560 [37] V.E. Pitzer, G.M. Leung, M. Lipsitch, Estimating variability in the transmission of severe
561 acute respiratory syndrome to household contacts in Hong Kong, China, American Journal of
562 Epidemiology. ?? (2007) ??.
- 563 [38] H. Nishiura, M. Schwehm, M. Kakehashi, M. Eichner, Transmission potential of primary
564 pneumonic plague: time inhomogeneous evaluation based on historical documents of the
565 transmission network, J. Epidemiol. Community Health. 60 (2006) 640-645.

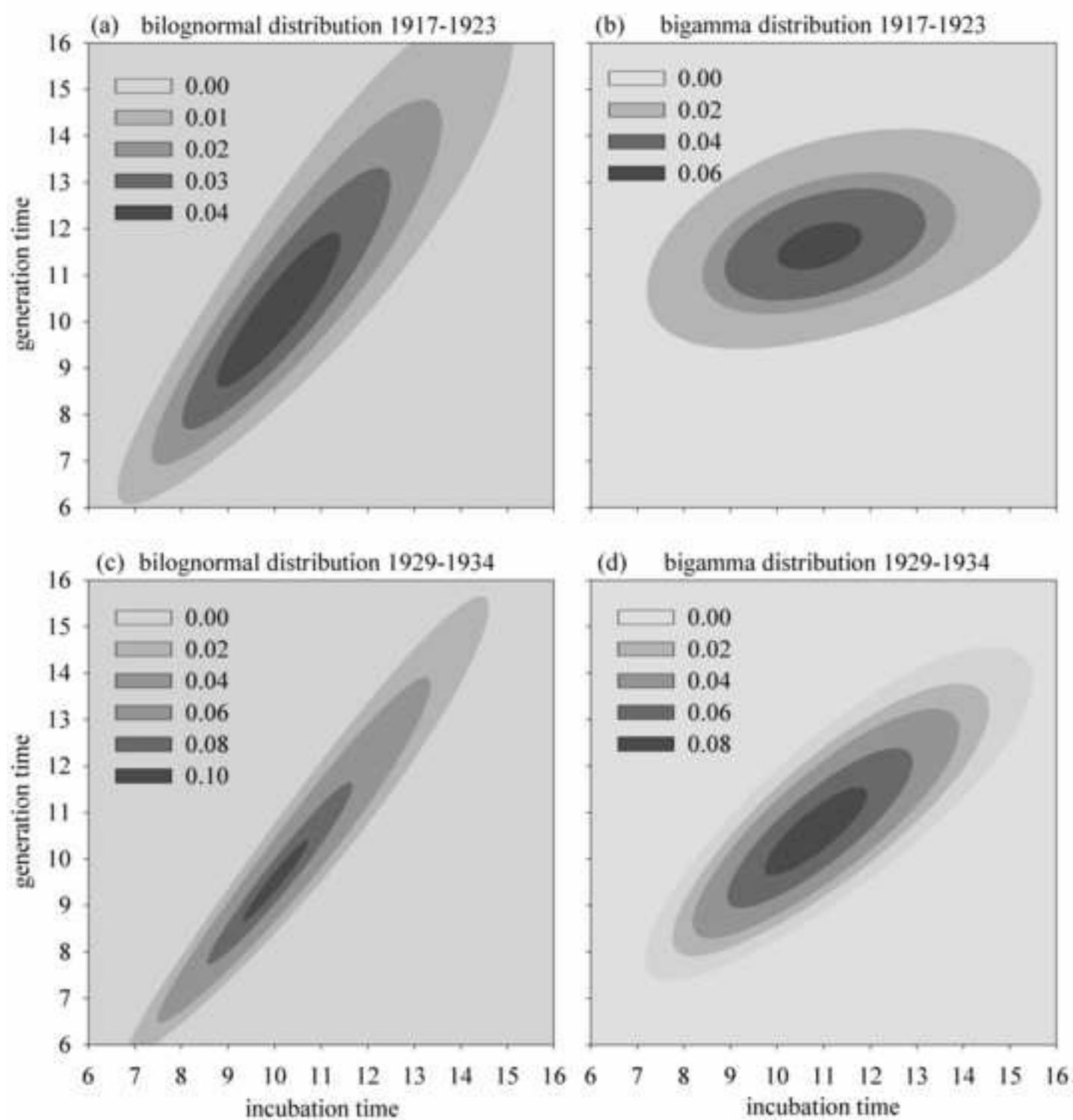
566







4. Figure 3



HIGHLIGHTS

- Household measles data of time intervals between symptom onset of two cases were analysed
- Correlation between incubation time and generation time were explicitly modelled
- Measles incubation time and generation time appear positively correlated
- The positive correlation could affect the effectiveness of case isolation and quarantine