

## Supplementary materials

Fine-scale family structure shapes influenza transmission risk in households: insights from primary schools in Matsumoto city, 2014/15.

Akira Endo, Mitsuo Uchida, Adam Kucharski, Sebastian Funk

### 1. Model selection

#### Model selection on the complexity

In the first round of model selection, we compared models with different complexity. Our household transmission model was mainly characterised by two components, the effective household contact  $\eta_{kl} = \beta \frac{c_{kl}}{c_k^\gamma}$  and the risk of external infection  $\varepsilon_k$ . Models corresponding to all possible combinations of assumptions were compared based on the Widely-applicable Bayesian information criterion (WBIC) (1). WBIC has the same scale as the Bayesian information criterion. A difference of 2 in WBIC is considered as an indication of statistical significance, while a difference greater than 5 is deemed as strong support. Table S1 compares the candidate models and their WBIC.

The parameters  $c_{kl}$  and  $\varepsilon_k$  were estimated as a single value  $c_{kl} = c$  and  $\varepsilon_k = \varepsilon$  under “Homogeneous”/“Uniform” assumptions, respectively. We fixed  $\gamma$  at 0 in “DD” (density-dependent) models and 1 in “FD” (frequency-dependent), and freely estimated in “IM” (intermediate) models. In “Single parent-Y” models, fathers and mothers who do not live with a spouse were classified as an additional type “single parent” (thus the number of types was 6 in these models). The best model (Model 12) was selected with very strong support:  $\Delta$ WBIC from the second-best model was 16.9.

#### Model selection on the contact pattern matrix

After selecting the model complexity, we further tried to explore different contact pattern matrices  $c_{kl}$ . Let the rows and columns of  $c_{kl}$  correspond to (Student, Sibling, Father, Mother, Other). Five parameters ( $c_{CC}, c_{FC}, c_{MC}, c_{OC}, c_{AA}$ ) being denoted by numbers 1 to 5, the contact pattern matrix  $c_{kl}$  in the previous model selection had the following structure:

$$c_{kl} = \begin{bmatrix} 1 & 1 & 2 & 3 & 4 \\ 1 & 1 & 2 & 3 & 4 \\ 2 & 2 & 5 & 5 & 5 \\ 3 & 3 & 5 & 5 & 5 \\ 4 & 4 & 5 & 5 & 5 \end{bmatrix} \quad (\text{S1})$$

Note that the diagonal elements for student, father and mother were displayed only for completeness and not used in the analysis (households in our dataset did not contain more than one students/fathers/mothers). Parameter estimates in Model 12 are shown in Table S2. In this contact pattern matrix, as all adults are assumed to share the same contact intensity. Meanwhile, the estimates of  $c_{FC}$  and  $c_{OC}$  are relatively similar. We explored variant models that further stratify  $c_{AA}$  while

$c_{FC}$  and  $c_{OC}$  are equated to keep the number of parameters unchanged (=5).

We considered the following submodels: Model 12a (intense contact within couples), Model 12b (mother acting as a hub) and Model 12c (generation-assortative).

$$\begin{aligned}
 c_{kl}(\text{Model 12a}) &= \begin{bmatrix} 1 & 1 & 2 & 3 & 2 \\ 1 & 1 & 2 & 3 & 2 \\ 2 & 2 & 4 & 4 & 5 \\ 3 & 3 & 4 & 4 & 5 \\ 2 & 2 & 5 & 5 & 5 \end{bmatrix}, \\
 c_{kl}(\text{Model 12b}) &= \begin{bmatrix} 1 & 1 & 2 & 3 & 2 \\ 1 & 1 & 2 & 3 & 2 \\ 2 & 2 & 5 & 4 & 5 \\ 3 & 3 & 4 & 4 & 4 \\ 2 & 2 & 5 & 4 & 5 \end{bmatrix}, \\
 c_{kl}(\text{Model 12c}) &= \begin{bmatrix} 1 & 1 & 5 & 3 & 5 \\ 1 & 1 & 5 & 3 & 5 \\ 5 & 5 & 2 & 2 & 5 \\ 3 & 3 & 2 & 2 & 5 \\ 5 & 5 & 5 & 5 & 4 \end{bmatrix},
 \end{aligned} \tag{S2}$$

Estimated contact pattern matrices are shown in Tables S3-S5. Models 12a and 12c had much better WBIC than Model 12 ( $\Delta\text{WBIC} = -14.4$  and  $\Delta\text{WBIC} = -17.3$ , respectively), while that of Model 12b was slightly worse than Model 12. Of the two models exhibiting improved WBICs, Model 12c was selected with a significant WBIC difference of 2.9. Parameter estimates other than  $c_{kl}$  did not vary between compared models to the first significant figure.

### Selection of the scaling factor

In our baseline model, the total amount of contacts  $C_k = \sum_l c_{kl}$  was used to scale the effective household contact (i.e.,  $\eta_{kl} \propto C_k^{-\gamma}$ ) to reflect heterogeneous contact patterns. On the other hand, previous modelling studies often used household size  $N$  in place of  $C_k$  (2–5). Although  $C_k$  and  $N$  are correlated ( $C_k$  and  $N-1$  coincide in homogeneous settings) and may work as a good proxy with each other, we considered comparison between these two approaches to be of interest. We tested a variant of Model 12c where  $C_k$  is replaced with  $N$  (i.e.,  $\eta_{kl} \propto N^{-\gamma}$ ), but the model performance was significantly worsened ( $\Delta\text{WBIC} = 8.8$ ). The estimated value of gamma did not change ( $\gamma = 0.52$ ; CrI: 0.34-0.75). The use of the total amount of contacts is preferred to household size as a scaling factor for the within-household transmission, and even when household size is used as variable, the semi-density-dependent model may still be applicable.

Table S1. WBIC of models with different sets of assumptions.

Model ID	$c_{kl}$	$\varepsilon_k$	$\gamma$	Single parent	WBIC	$\Delta$ WBIC
1	Hom	Unif	DD	N	33269.16	2134.96
2	Het	Unif	DD	N	33054.70	1920.50
3	Hom	Unif	FD	N	33259.36	2125.16
4	Het	Unif	FD	N	32731.10	1596.90
5	Hom	Unif	IM	N	33243.32	2109.12
6	Het	Unif	IM	N	32277.92	1143.72
7	Hom	Strat	DD	N	31215.16	80.96
8	Het	Strat	DD	N	31151.08	16.88
9	Hom	Strat	FD	N	31205.44	71.24
10	Het	Strat	FD	N	31150.78	16.58
11	Hom	Strat	IM	N	31186.64	52.44
12	Het	Strat	IM	N	31134.20	0
13	Hom	Unif	DD	Y	33267.72	2133.52
14	Het	Unif	DD	Y	33061.14	1926.94
15	Hom	Unif	FD	Y	33256.72	2122.52
16	Het	Unif	FD	Y	32752.26	1618.06
17	Hom	Unif	IM	Y	33241.46	2107.26
18	Het	Unif	IM	Y	32182.18	1047.98
19	Hom	Strat	DD	Y	31223.68	89.48
20	Het	Strat	DD	Y	31167.16	32.96
21	Hom	Strat	FD	Y	31212.52	78.32
22	Het	Strat	FD	Y	31168.08	33.88
23	Hom	Strat	IM	Y	31194.82	60.62
24	Het	Strat	IM	Y	31151.98	17.78

Hom: homogeneous mixing, Het: Heterogeneous mixing

Unif: uniform risk of external infection, Strat: stratified risk of external infection

DD: density-dependent, FD: frequency-dependent, IM: intermediate

Single-Parent: whether the “single parent” category has a unique parameter. Y=Yes, N=No.

WBIC: Widely-applicable Bayesian information criterion,  $\Delta$ WBIC: WBIC difference from the best model

Table S2. Estimated contact pattern matrix ( $c_{kl}$ ) in Model 12.

	Student	Sibling	Father	Mother	Other
Student	1.28	0.54	1.40	0.45	
Sibling					
Father	0.54	1			
Mother	1.40				
Other	0.45				

WBIC = 31134.20;  $\Delta$ WBIC = 0 (baseline)

Table S3. Estimated contact pattern matrix ( $c_{kl}$ ) in Model 12a.

	Student	Sibling	Father	Mother	Other
Student	0.97	0.39	1.09	0.39	
Sibling					
Father	0.39	1		0.39	
Mother	1.09				
Other	0.39	0.39	1		

WBIC = 31119.78;  $\Delta$ WBIC = -14.42

Table S4. Estimated contact pattern matrix ( $c_{kl}$ ) in Model 12b.

	Student	Sibling	Father	Mother	Other
Student	1.25	0.49	1.37	0.49	
Sibling					
Father	0.49	1.01	1		1.01
Mother	1.37				
Other	0.49	1.01	1.01		

WBIC = 31134.72;  $\Delta$ WBIC = 0.54

Table S5. Estimated contact pattern matrix ( $c_{kl}$ ) in Model 12c.

	Student	Sibling	Father	Mother	Other
Student	1.04	0.43	1.16		
Sibling					0.43
Father	0.43	1			
Mother	1.16				
Other	0.43			1.97	

WBIC = 31116.88;  $\Delta$ WBIC = -17.32 (best model)

## 2. Source-stratified risk of infection and risk attributable to the introduction of influenza into a household

We quantified the risk of infection attributable to external and within-household infection from the parameter estimates. Three family compositions were selected as model cases: (a) “nuclear family”: father, mother and two children, (b) “many-siblings family”: father, mother and four children, and (c) “three-generation family”: father, mother, two children and two grandparents. We assumed that one of the children in each model case households was “student”, and the others were “siblings”. The overall risk of infection for type  $k$  individual is given by

$$r_k = \sum_{\mathbf{n}} \frac{n_k}{N_k} \pi(\mathbf{n}; \mathbf{N}, \boldsymbol{\varepsilon}, H), \quad (\text{S3})$$

(the sum is taken for all possible  $\mathbf{n}$ ), and  $r_k - \varepsilon_k$  corresponds to the additional infection risk due to within-household transmission.

We also compared the risk of infection after the introduction of influenza into households with the initial overall risk. We defined post-introduction risk as the conditional probability that an individual experience infection by the end of the season, given that one index case is already observed in the same family. Here, for simplicity, we limited the analysis to introductions by primary school students (i.e., individual type “student”) only.

Suppose that  $k=1$  corresponds to the type “student”. Post-introduction risk obtained by modifying the formula for  $r_k$  as

$$r_k^{\text{pos}} = \sum_{\{\mathbf{n}|n_1=0\}} \frac{n_k}{N_k} \cdot \frac{\pi(\mathbf{n}; \mathbf{N}, \boldsymbol{\varepsilon} + \mathbf{H}_1, H)}{S_1(\mathbf{n}, \boldsymbol{\varepsilon})}. \quad (\text{S4})$$

$\mathbf{H}_1$  is the (additional) force of infection arising from the infected student, i.e.,  $(\mathbf{H}_1)_k = \frac{c_{k1}}{c_k^y}$ . The sum is taken for all possible  $\mathbf{n}$  whose first component  $n_1 = 0$  (because the force of infection from the student is incorporated in  $\mathbf{H}_1$ ). Note that by dividing  $h(\mathbf{n}; \mathbf{N}, \boldsymbol{\varepsilon} + \mathbf{H}_1, H)$  by  $S_1(\mathbf{n}, \boldsymbol{\varepsilon})$ , we can yield the conditional probability that  $\mathbf{n}$  individuals (other than the “student”) are infected given the presence of the force of infection  $\mathbf{H}_1$ .

## 3. Sensitivity analysis

### Procedures for the sensitivity analysis

#### (i) Ascertainment bias

In order to account for potential ascertainment bias, we incorporated reporting probabilities into the model. We assumed that infections are reported with a certain probability  $p_k$ . Epidemiological properties, such as infectiousness, were assumed to be identical between reported and unreported cases. The likelihood of observing a household final size outcome  $(\mathbf{n}; \mathbf{N})$  given the reporting probability vector  $\mathbf{p}$  is obtained by using the binomial distribution:

$$L(\boldsymbol{\varepsilon}, H, \mathbf{p}; (\mathbf{n}; \mathbf{N})) = \sum_{\mathbf{n}' \geq \mathbf{n}} h(\mathbf{n}'; \mathbf{N}, \boldsymbol{\varepsilon}, H) \prod_k \text{Bin}(n_k; n'_k, p_k). \quad (\text{S5})$$

The sum  $\sum_{\mathbf{n}' \geq \mathbf{n}}$  is taken for all vector  $\mathbf{n}'$  satisfying  $n_k \leq n'_k \leq N_k$  ( $\forall k$ ).

In this sensitivity analysis, we assumed that the reporting probability  $p$  for children (“student” and “sibling”) is 0.8. The reporting probability for adults was varied from 0.5 to 0.8.

### (ii) Different susceptibility in children

Susceptibility to influenza infection was differentiated between children and adults. Let  $\sigma$  be the susceptibility of children relative to that of adults. The effect of  $\sigma$  was employed in the model by differentiating the transmissibility  $\beta$  as

$$\beta_k = \begin{cases} \beta\sigma & (k = \text{"Student"}, \text{"Sibling"}) \\ \beta & (\text{otherwise}) \end{cases}. \quad (\text{S6})$$

Five different values of  $\sigma$ : 0.75, 1.25, 1.5, 1.75, 2.0 were tested.  $\sigma = 0.75$  corresponds to the assumption that children may have less risk of infection per exposure (e.g., due to potentially high vaccination coverage). The value of  $\sigma$  greater than 1 reflects the assumption that children are more vulnerable than adults.

### (iii) Multiple counting of households

We identified all the possible combinations of respondents who might be from the same household by the following process. (1) Respondents were classified by their school and family composition. We assumed that siblings usually go to the same primary school. (2) Data were matched up, and consistency was checked between the sex and grade of the respondent and the reported composition of siblings. For instance, a second-grade boy who has no older brother should not be from the same household as a fourth-grade girl who has an older brother (the girl’s older brother should also be the boy’s older brother). Here, we assumed that siblings should be in different grades, and neglected the possibility of twins or siblings in the same grade. (3) Individuals potentially from the same household were grouped together. Combination of grouping was chosen so that as many individuals as possible are grouped together in total. Individuals in each matched group were assumed to be from the same household, and their data were integrated to represent one household data. Respondents were classified as “students”, and siblings who were not found in the dataset was classified as “siblings”. Because sex, school and grade were not used in the parameter estimation, individual-level details of the grouping arrangement (who was grouped with whom) did not affect the subsequent analysis. Through this whole process, 1,294 individuals identified as candidates potentially from multiple-counted households were processed, reducing the number of households from 10,486 to 9,763 (-6.9%). Note that this is an extreme case where as many consistent siblings as possible are grouped together, and that the reality may lie between the two extremes (no-grouping and maximum-grouping).

(iv) Case censoring

Participants reported in the survey the total number of siblings, siblings in four categories (older brother, older sister, younger brother or younger sister) and whether siblings in each category had influenza. Let  $\mathbf{M} = (M_1, M_2, M_3, M_4)$  and  $\mathbf{m} = (m_1, m_2, m_3, m_4)$  be the true composition of siblings and the number of siblings with an influenza episode in each category (1: older brother; 2: older sister, 3: younger brother; 4: younger sister), respectively. Due to the questions in the survey, the dataset did not include either  $\mathbf{M}$  or  $\mathbf{m}$ . Instead, we have  $M = \sum_i M_i$  and censored sibling data  $\mathbf{M}'$  ( $M'_i = \min(M_i, 1)$ ) and  $\mathbf{m}'$  ( $m'_i = \min(m_i, 1)$ ), as the questions on sibling categories were yes-no questions.

We constructed a modified likelihood function to address this censoring issue. The basic idea was to generate all possible patterns of  $\mathbf{M}$  and  $\mathbf{m}$  that are consistent with the observation and aggregate the corresponding probabilities to obtain the likelihood for the censored data. First, we defined a conditional probability  $\pi(\mathbf{M}; M)$ , the probability that the true sibling composition is  $\mathbf{M}$  given  $M$ . Assuming that the probability of being the  $n$ -th child in given  $M$  siblings is equally  $\frac{1}{M+1}$  and that the sex of a child is evenly distributed, we get

$$\pi(\mathbf{M}; M) = \frac{1}{(M+1) \cdot 2^M} \binom{M_1 + M_2}{M_1} \binom{M_3 + M_4}{M_3} \sigma(\mathbf{M}, M), \quad (\text{S7})$$

where  $\sigma(\mathbf{M}, M)$  is an indicator function that takes 1 if  $\mathbf{M}$  is consistent with  $M$  (i.e.,  $\sum_i M_i = M$ ), and 0 otherwise.

Let  $\pi(\mathbf{m}; \mathbf{M}, \varphi)$  be the probability of observing a sibling outcome pattern  $\mathbf{m}$  given  $\mathbf{M}$ . This is also conditional to the existence of other family members and their outcomes, and those conditions are represented by  $\varphi$ .

Using  $\pi(\mathbf{M}; M)$  and  $\pi(\mathbf{m}; \mathbf{M}, \varphi)$ , the likelihood of observing  $\{\mathbf{M}', \mathbf{m}'\}$  given  $M$  and  $\varphi$  is

$$l(\mathbf{M}', \mathbf{m}'; M, \varphi) = \sum_{\mathbf{M}} \pi(\mathbf{M}; M) \sigma(\mathbf{M}, \mathbf{M}') \sum_{\mathbf{m}} \pi(\mathbf{m}; \mathbf{M}, \varphi) \sigma(\mathbf{m}, \mathbf{m}'), \quad (\text{S8})$$

where  $\sigma(\mathbf{M}, \mathbf{M}')$  and  $\sigma(\mathbf{m}, \mathbf{m}')$  are indicator functions checking if  $\mathbf{m}$  and  $\mathbf{M}$  are consistent with the observation.

Since we assume all siblings exhibit identical epidemiological behaviour, considering the effect of loss of distinguishability,  $\pi(\mathbf{m}; \mathbf{M}, \varphi)$  is substituted with

$$\pi(\mathbf{m}; \mathbf{M}, \varphi) = \frac{\prod_i \binom{M_i}{m_i}}{\binom{M}{m}} \pi(m; M, \varphi), \quad (\text{S9})$$

where  $m = \sum_i m_i$ .  $\pi(m; M, \varphi)$  is equivalent to  $\pi$  in equation (1) in the main text, and thereby we get the likelihood accounting for possible case censoring in siblings.

### **Results of the sensitivity analysis**

The estimates from the sensitivity analysis were compared in Figure S1. Equally lowering the reporting probability for children and adults slightly increased some of the parameters while the overall relative magnitude was almost conserved. When the reporting probability for adults was set lower than children, parameters which involve adults increased and those involving children decreased (Figures S1A and S1B). Increasing the relative susceptibility in children resulted in lower child-involved contact intensities (Figures S1C and S1D). Multiple counting of data reported by students from the same household did not seem to have affected the result, but some changes were caused by addressing censored cases in siblings, which may be resulted from the possibility of unobserved sibling cases (Figures S1E and S1F). Except that the contact intensity between children was substantially lowered by either underreporting of adults or high susceptibility in children, the relative trend remained almost similar throughout our sensitivity analysis. Especially, the risk of external infection in children and the contact intensity between children and adults remained at a sufficient level, such that the secondary transmission from children is still of paramount importance. The exponent parameter  $\gamma$  was stable throughout the sensitivity analysis (median within  $0.50 \pm 0.02$ ), except that it was slightly higher (0.59; CrI: 0.39-0.79) when the case censoring (iv) was considered.



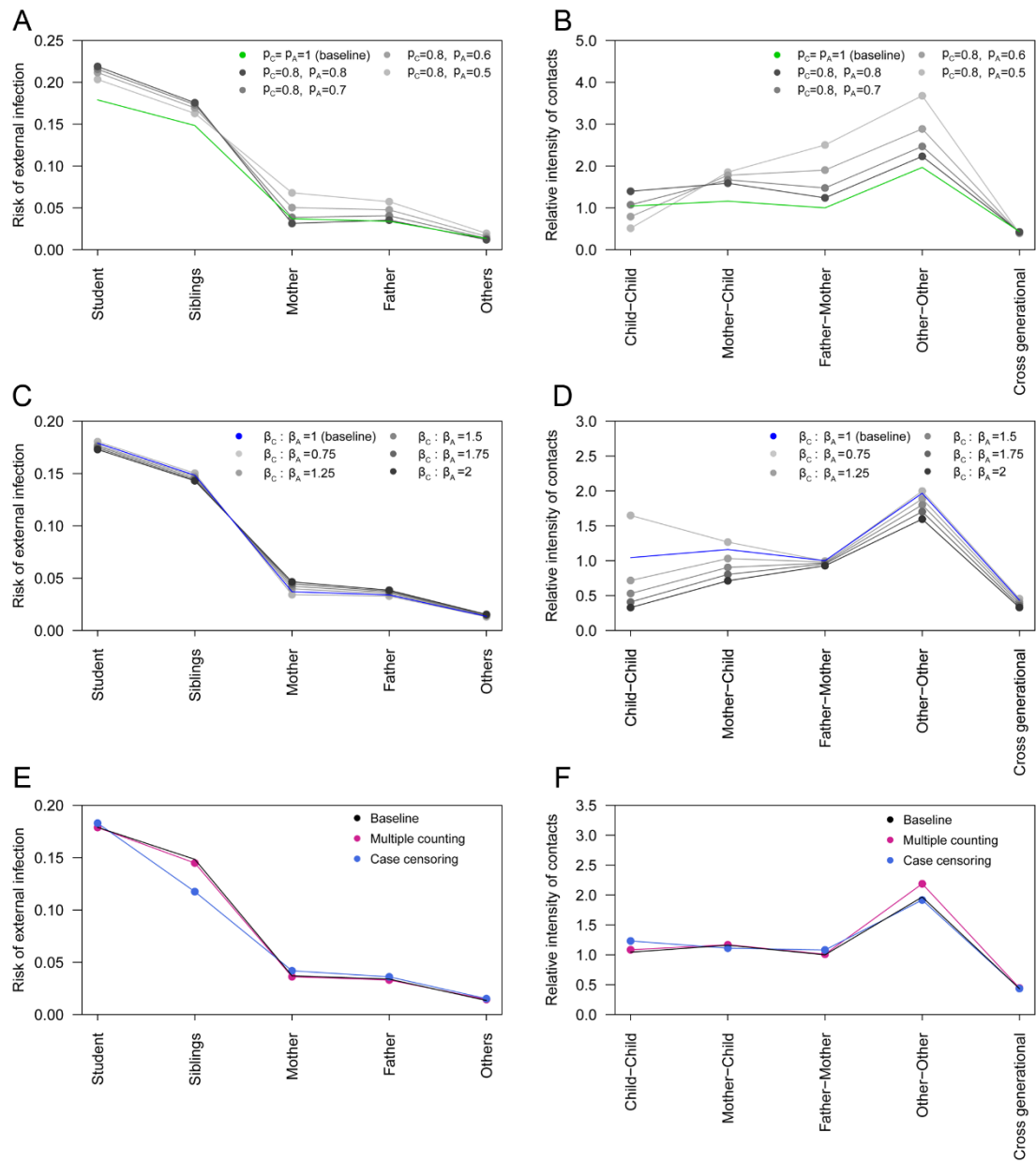


Figure S1. Parameter estimates from the sensitivity analysis. The estimated risk of external infection and relative intensity of household contacts are compared with the baseline estimates. The relative intensity of contacts in the figures is multiplied by the relative change in the estimated transmissibility parameter  $\beta$  for comparability.

(A), (B) Various reporting probabilities in children ( $p_C$ ) and adults ( $p_A$ ).

(C), (D) Various ratios between susceptibility in children ( $\beta_C$ ) and adults ( $\beta_A$ ).

(E), (F) Estimates from the modified dataset addressing multiple counting of households and censoring of cases in siblings.

#### 4. Model fit

To evaluate the goodness-of-fit of our model, the model prediction was compared with the observed data. Let  $\hat{\theta}$  be the set of median parameter estimates.  $\pi(\mathbf{n}; \mathbf{N}, \hat{\theta})$ , the probability of observing outcome pattern  $\mathbf{n}$  given household composition  $\mathbf{N}$ , is obtained from Equation (1) in the main text. Assuming that the distribution of  $\mathbf{N}$  in dataset  $D$  is given as observed ( $\pi_D(\mathbf{N})$ ), the predictive distribution of the outcome patterns  $(\mathbf{N}_i, \mathbf{n}_i)$  (approximated by the point estimate  $\hat{\theta}$ ) is

$$\pi(\mathbf{N}_i, \mathbf{n}_i; \hat{\theta}) = \pi(\mathbf{n}_i; \mathbf{N}_i, \hat{\theta})\pi_D(\mathbf{N}_i), \quad (\text{S10})$$

Figure S2 compares the predictive distribution with the actual frequency in the dataset. The 95% intervals are approximated by the 95% quantiles of a binomial distribution

$$F_D(\mathbf{N}, \mathbf{n}) \sim \text{Binom}\left(F_D(\mathbf{N}), \pi(\mathbf{n}; \mathbf{N}, \hat{\theta})\right), \quad (\text{S11})$$

where  $F_D$  is the frequency in data  $D$  of size  $S_D$ . The predicted and observed frequency show good accordance despite the relatively modest parameter space dimension (=11). The similarity between the two distributions are also supported by the empirical Kullback-Leibler divergence of 0.05, where

$$\widehat{\text{KL}} = \sum_d \frac{F_D(\mathbf{N}, \mathbf{n})}{S_D} \cdot \log\left(\frac{F_D(\mathbf{N}, \mathbf{n})}{\pi(\mathbf{n}; \mathbf{N}, \hat{\theta})F_D(\mathbf{N})}\right). \quad (\text{S12})$$

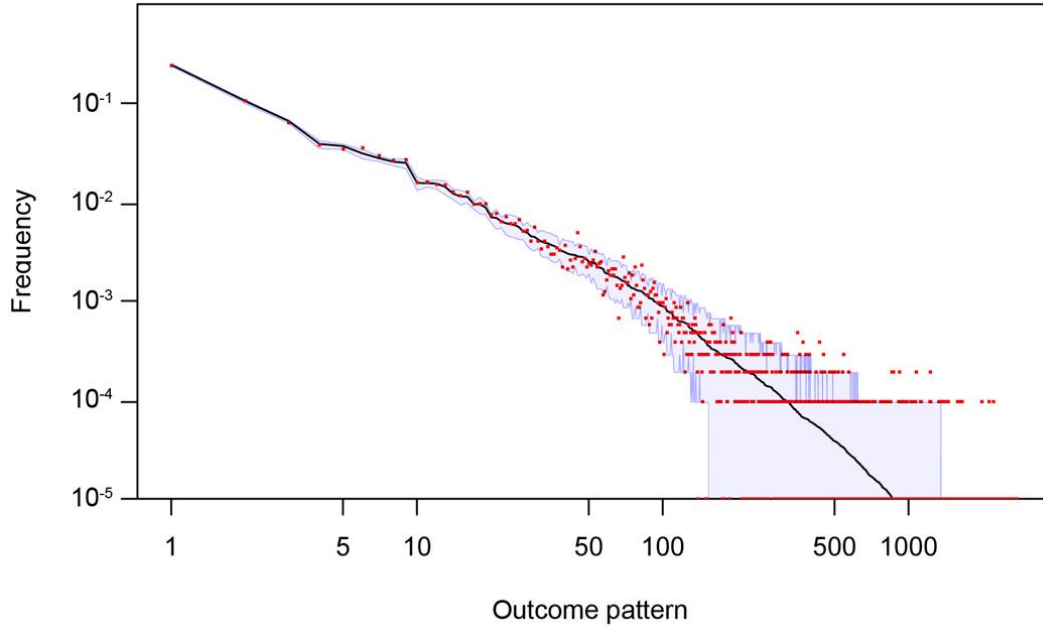


Figure S2. Comparison between the predicted and observed household final outcomes.

Red dots correspond to the observed relative frequency of data (household compositions and final outcomes of the household members), where the x-axis denotes the numbering of outcome patterns  $(\mathbf{N}, \mathbf{n})$ . With the sample size of  $\sim 10,000$ ,  $10^{-4}$  on the y-axis denotes frequency 1; dots for frequency 0 are shown on the x-axis. The black line indicates the probability of observation predicted by the model, and the shaded area shows 95% intervals. Both x- and y-axes are on a logarithmic scale.

We also compared the predicted and observed distributions of the final attack size (the total number of household cases during the season) for specific compositions in Figure S3. The observed distribution was right-skewed from the “binomial scenario”, where within-household transmission is not present and individuals are assumed to be exposed to the external risk of infection only.

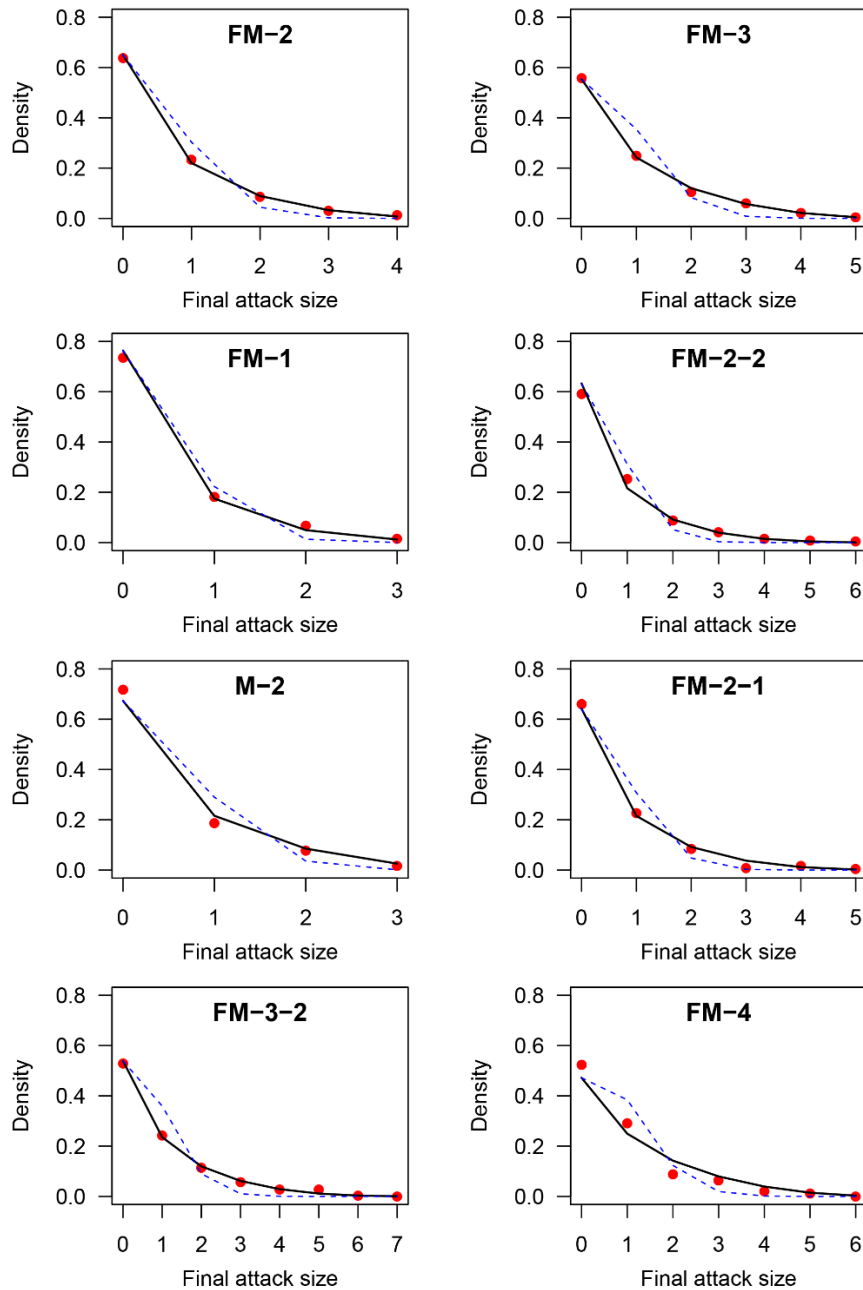


Figure S3. The observed and predicted final attack size distributions.

Red dots and black lines denote the observed and predicted relative frequencies. Blue lines represent “binomial scenario”, where within-household transmission is not present. Eight major household compositions (accounting for 84% of the total households in the dataset) are shown.