Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Cauchemez S, Donnelly CA, Reed C, et al. Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. N Engl J Med 2009;361:2619-27.

Supplementary Material

Transmission of novel Influenza A(H1N1) virus in households in the USA

Simon Cauchemez[1], Christl A. Donnelly[1], Carrie Reed[2], Azra C. Ghani[1], Christophe Fraser[1], Charlotte Kent[2], Lyn Finelli[2], Neil M. Ferguson[1]

[1]: MRC Centre for Outbreak Analysis and Modelling, Imperial College London, London

[2]: CDC, Atlanta

Contents

1 Transmission model

Here, we build a transmission model that accounts for the fact that sick contacts may be infected by other contacts in the wider community (community infections) or by other sick household members (tertiary infections) rather than by the index case. Inference is then made conditional on what is observed in the household on day 0.

Consider a household of size *n*. The age of household member *i* (*i=1,…,n*) is denoted a_i. Cough in household member *i* is denoted by an indicator variable $s_i^C = 1$; $s_i^C = 0$ otherwise. Similarly, fever, runny nose, sore throat and diarrhea are indicated by indicator variables s_i^F, s_i^R, s_i^S and s_i^D, respectively. The vector of symptoms is denoted $s_i = \left\{ s_i^C, s_i^F, s_i^R, s_i^S, s_i^D \right\}$.

We denote d_i the day of symptom onset for household member *i*. By notation $d_i=\infty$ if no symptom is observed in the household member during the study period and day 0 corresponds to symptom onset in the index case of the household. The data is censored at day *T*=7 days.

Parameters of the model are then estimated in a Markov chain Monte Carlo (MCMC) Bayesian framework [1-2].

1.1 Person-to-person hazard of infection in the household

Consider household member *i* with onset on day d_i (not necessarily the index case of the household). The hazard that household member *i* infects household member *j* with onset on day *d* (*d*>d_i) is modeled as follows:

$$P_{i\to j}^{H}\left(d\right) = g\left(n\right)\mu_{inf}\left(a_i, s_i\right)\mu_{sus}\left(a_j\right) f\left(d - d_i\right) \qquad (1)$$

where

- Function *g(n)* characterizes the dependency between the person-to-person hazard of transmission and the household size *n*. In the baseline scenario presented in the text, we consider a non-parametric model where 1 parameter is estimated per household size: $g\left(n\right) = \beta_n$ for *n*=2,…,6. In the Sensitivity analysis, four alternative model variants are considered. See section 3.2 for details. Figure 1A of the paper plots g(n)/g(2) for the non-parametric model.

- $\mu_{inf}\left(a_i, s_i\right)$, the relative infectivity of household member i, may depend on the age and the symptoms of the case. For example, in the model where age is the only predictor for infectivity, the term simplifies as:

$$\mu_{inf}\left(a_i, s_i\right) = \begin{cases} r_{0-18} & \text{if } 0 \le a_i \le 18 \\ 1 & \text{if } 19 \le a_i \le 50 \\ r_{+51} & \text{if } 51 \le a_i \end{cases}$$

If more than one predictor is considered, a multiplicative model is assumed. In order to improve stability of estimates, we selected reference groups with the largest number of individuals, that is: '19-50 yr old' for age, and for the symptoms 'cough', 'fever', 'runny nose', 'sore throat' and 'no diarrhea'. Results are presented in section 3.7. A sequence of likelihood ratio tests is performed to select predictors of infectivity.

- $\mu_{sus}\left(a_j\right)$, the relative susceptibility of household member *j*, may depend on the age of the individual. The way relative susceptibility was modeled was similar to that of relative infectivity (see above). A sequence of likelihood ratio tests is performed to determine if age is a predictor of susceptibility.

- Function *f(.)* is the probabilistic density of the serial interval, defined as the distribution of the interval between symptom onset in a case and symptom onset in the cases that they generate in household contacts in the limit of infinitesimal secondary attack rate (i.e. when depletion of susceptible does not have an impact). Here, we use a discretized Weibull distribution, with density

$$f\left(u\right)=\begin{cases} 0 & \text{if } u<L \\ exp\left\{-\left(\frac{u-L}{q}\right)^k\right\}-exp\left\{-\left(\frac{u+1-L}{q}\right)^k\right\} & \text{otherwise} \end{cases} \quad (2)$$

We assume that transmission is possible only if the delay between symptoms onset of a case and of the household members they infect is ≥1 day (*L*=1 day). In practice, it means that if two household members A and B both have onset on day *d*, we make the assumption that it is not possible for B to have been infected by A.
In the sensitivity analysis, we also investigate if the serial interval depends on the age of the infector (see section 3.4). We report the mean and standard deviation of the discretized distribution.
The reason for using such a seemingly abstract definition of the serial interval (rather than simply characterise the empirical distribution) is generalisability. Using our estimated serial interval distribution and the transmission parameters we have estimated for households, we can reconstruct the empirical household serial interval distribution. But in addition, past experience has shown that we can expect the 'infinitesimal hazard' serial interval distribution to be more applicable to transmission outside the household than the simple empirical distribution obtained from household data.
As illustrated in [1], we could have used more explicit models of how infectivity varies over time following infection. However, such approach would have required to make an assumption about the incubation period of novel A(H1N1) influenza virus and results would have been conditional on this assumption (for which very limited data was available at the time of the analysis). We felt that estimating the onset-to-onset serial interval was a more transparent way to describe the timing of events in the household. From the onset-to-onset serial interval distribution presented in Figure 2

and an assumption on the distribution of the incubation period, it is possible to reconstruct the infectivity profile according to time since onset. Figure 2 indicates that for an individual infected by a household member, it is expected that there is 91% probability that onset occurs less than 4 days after onset in their infector. Given that transmission occurs prior to onset in the individual, this means that more than 91% of household transmissions are expected to occur less than 4 days after onset in the infector.

1.2 Community risk of infection

We assume that the hazard that household member *j* with onset on day *d* is infected in the community is constant during the duration of the follow-up:

$$P_j^C\left(d\right)=\alpha$$

In the sensitivity analysis, we investigate whether the community risk of infection is age dependent and whether estimates of other parameters are robust to an exponential increase of the community risk as the disease spreads in the country (see section 3.3).

Although the transmission model provides estimates of the daily rate of infection from the community, those rates are estimated on a subset of households that were exposed to novel Influenza A(H1N1) virus. So, those rates cannot be used to be representative of the overall level of exposure of communities in the US to novel Influenza A(H1N1) during the early epidemic.

1.3 Hazard of symptom onset in a household member

The hazard that household member *j* has symptom onset on day *d* is the sum of the within-household transmission hazards as well as the community transmission hazard:

$$\lambda_j\left(d\right)=\sum_{i=\{1,..,n;i\neq j\}}P_{i\to j}^H\left(d\right)+P_j^C\left(d\right)$$

2 Inference

Parameter vector θ of the transmission model is estimated in a likelihood-based Bayesian inferential framework.

2.1 Likelihood function

Inference is made conditional on what is observed in the household on day 0. Given the sequence of symptom onset observed up to day *d* (*d*>0), the probability that household member *j* has symptom onset on day *d* is:

$$P\left(d_j=d\mid\theta\right)=\left(1-\exp\left\{-\lambda_j\left(d\right)\right\}\right).\exp\left\{-\sum_{0<u<d}\lambda_j\left(u\right)\right\}$$

The contribution to the likelihood of household member *j* who does not become a case during the study period is:

$$P\left(d_j = d \mid \theta\right) = \exp\left\{-\sum_{0<u\leq T} \lambda_j\left(u\right)\right\}$$

Index cases and co-primary cases do not contribute to the likelihood.

2.2 Priors

For parameters characterizing the relative infectivity of cases (e.g. “cough” relative to “no cough”) or the relative susceptibility of household members (e.g. “0-18yr” relative to “19-50yr”), a log-normal prior distribution with log-mean=0 and log-variance=3 was selected. This prior satisfies the invariance condition that the ratio (adult susceptibility/child susceptibility) has the same prior as the ratio (child susceptibility/adult susceptibility). In particular, it gives equal probabilities to the relative child susceptibility being larger or smaller than 1. For log-variance=3, [2.5%,97.5%] percentile of the prior distribution is [0.03,29.81]. In the sensitivity analysis, we explored the impact of a change in the prior on estimates (considering the log-normal distribution with log-variance 1, 2, 4 and 5) (section 3.6).

We used the prior information that less than 20% of transmission occur after day 14 (day 0 being defined as symptom onset in the case). This is supported by a large number of viral shedding and transmission studies[3]. This prior was used to ensure good convergence of the MCMC chain. Further explanations as well as a sensitivity analysis are presented in section 3.6.

All other parameters had priors that were flat between 0 and M, where M was an artificially large number.

2.3 Algorithms

The joint posterior distribution of the parameters is explored via MCMC sampling[1-2]. Parameters were updated with a Metropolis-Hastings algorithm[2]. The algorithm runs for 3,200,000 iterations with a burn-in of 30,000 iterations. One out of 40 iterations was recorded. Converge was visually assessed. The Deviance Information Criterion (DIC) for model comparison was computed[4].

A simulated annealing algorithm was also implemented to maximize the likelihood and compute the likelihood-ratio test statistic.

3 Model comparison and sensitivity analysis

To ensure that key findings were robust to modeling assumptions, we explored and compared a large number of model variants in a sensitivity analysis. The different assumptions that were tested are detailed below. In summary, 26 model variants (and combinations of those variants) were investigated / tested:

- 4 models for age-specific susceptibility
- 5 for the dependency household size - transmission rate;
- 3 for the risk of infection from outside (constant; exponential growth; different for children and adults);
- 2 for the serial interval (1 distribution; different for children and adults);

- 2 approaches to deal with missing onset dates;
- 3 priors on the serial interval;
- 3 priors on relative infectivity and relative susceptibility parameters.
- 2 case definitions;
- 2 datasets.

As detailed below, key findings were found to be robust to those modeling assumptions. Comparison of models was done with a likelihood ratio test for embedded models; and relying on the DIC criteria otherwise. We also tested adequacy with a simulation-based χ^2 test comparing observed and expected distributions of the number of cases per household size (which p-value is referred to as p-value Exp-Obs in the tables).

3.1 Models for susceptibility

We compared 4 models for susceptibility according to the age of the susceptibility: no age-specific susceptibility (A), child relative susceptibility (B), older adult relative susceptibility (C), child and older adult rel. sus (D; baseline). Results are summarized in table SM1. Models B and C had a substantially better fit than model A according to DIC; and they were significantly better than model A according to the likelihood ratio test ($p<0.001$ for model B and $p=0.005$ for model C). Including child and older adult relative susceptibility in the same model (D) improved the DIC fit further. The likelihood ratio test indicates that both parameters are significantly different than 1, with $p=0.005$ for child relative susceptibility and $p=0.031$ for older adult relative susceptibility.

Table SM1: *Comparison of models for age-specific susceptibility.*

	A- No age specific sus.	B- Child rel. sus.	C- Older adult rel. sus.	D- Child and older adult rel. sus. (baseline)
g(2)	0.25[0.1,0.48]	0.21[0.09,0.42]	0.26[0.11,0.48]	0.22[0.1,0.43]
g(3)	0.15[0.08,0.26]	0.12[0.05,0.21]	0.17[0.08,0.28]	0.13[0.06,0.24]
g(4)	0.06[0.03,0.11]	0.04[0.02,0.08]	0.06[0.03,0.11]	0.05[0.02,0.09]
g(5)	0.05[0.01,0.1]	0.03[0.01,0.07]	0.05[0.02,0.11]	0.04[0.01,0.08]
g(6)	0.02[0,0.08]	0.01[0,0.05]	0.02[0,0.07]	0.01[0,0.06]
mean SI	2.7[2.1,4.2]	2.6[2.2,3.5]	2.7[2.2,4.5]	2.6[2.1,3.4]
sd SI	1.3[0.9,3.6]	1.3[0.9,2.4]	1.3[0.9,3.9]	1.2[0.9,2.3]
com. risk (10^{-4})	44.4[5,94.9]	42.8[7.3,91.5]	44.6[13.3,90.6]	41[7.2,86]
child rel. sus.	1	2.22[1.19,4.31]	1	1.96[1.05,3.78]
older adult rel. sus.	1	1	0.13[0.01,0.74]	0.17[0.02,0.92]
max LL^	-304.33	-298.86	-300.42	-296.52
DIC	623.96	615.61	617.45	611.91
p-value Exp-Obs$	0.9803	0.9792	0.9845	0.985

^: Log-likelihood (LL).

$: from a simulation-based $\chi 2$ test comparing observed and expected distributions of the number of cases per household size.

3.2 Dependency between transmission rate and household size

Five model variants were explored for the function *g(n)* that characterizes the dependency between the person-to-person transmission hazard and household size:

- The non-parametric model (baseline scenario), with one parameter per household size: $g(n) = \beta_n$ for *n*=2,…,6.
- The constant model: $g(n)=\beta$.
- Three parametric forms, with δ>0:
  - Power model A: $g(n) = \beta.(3/n)^{\delta}$
  - Power model B: $g(n) = \beta.\{2/(n-1)\}^{\delta}$
  - Exponential model: $g(n) = \beta.\exp\{-\delta(n-3)\}$

Results are presented in Table SM2. Adequacy of the constant model was rejected on the basis of the DIC criteria (DIC difference relative to other models >13) as well as the simulated χ^2 test comparing observed and expected distributions in total number of cases per household size (p<0.009). Furthermore, the assumption that the transmission hazard was independent of household size was formally rejected by the likelihood ratio test (p<0.001). Adequacy of the other 4 models as measured by the comparison Expected-Observed was good. A priori, for the parametric models, we assumed that parameter δ was larger than or equal to 0, excluding the scenario where the person-to-person transmission hazard increases with household size. Taking a wider prior for δ (e.g. uniform in [-10,10]) would not

change the posterior distribution of the parameter since there was very strong evidence in the data that δ was larger than 0 (based on the comparison of the constant model $\delta=0$ with other models).

Here, we observe a relatively sharp reduction in secondary attack rates between households of size 2 to 4 after what secondary attack rates are approximately constant (Figure 1A). This contrasts with observations made in a French household study on seasonal influenza[1], where the secondary attack rate was found to be approximately constant with household size. Those differences are intriguing and highlight that the sociological, environmental and biological mechanisms available to explain this dependency are still limited. For example, a standard model that provides a good fit to the French seasonal data is that people simply "share" their time between household members[1]. It is possible that the emergence of novel Influenza A(H1N1) pandemic triggered important but unobserved behavioral changes that affected more households of large size than others (for example because those household have higher proportions/numbers of children). Further research is needed to investigate those issues.

Table SM2: *Comparison of the models for the dependency of household transmission hazards with household size.*

	Non-parametric model (Baseline)	**Constant model**	**Power model A**	**Power model B**	**Exponential model**
coeff δ	-	-	2.51[1.46,3.77]	1.63[0.94,2.44]	0.81[0.46,1.27]
β	-	0.05[0.02,0.09]	0.08[0.04,0.14]	0.07[0.03,0.13]	0.09[0.04,0.15]
g(2)	0.22[0.1,0.43]	0.05[0.02,0.09]	0.23[0.11,0.41]	0.23[0.1,0.42]	0.2[0.09,0.37]
g(3)	0.13[0.06,0.24]	0.05[0.02,0.09]	0.08[0.04,0.14]	0.07[0.03,0.13]	0.09[0.04,0.15]
g(4)	0.05[0.02,0.09]	0.05[0.02,0.09]	0.04[0.02,0.07]	0.04[0.01,0.07]	0.04[0.01,0.08]
g(5)	0.04[0.01,0.08]	0.05[0.02,0.09]	0.02[0.01,0.05]	0.02[0.01,0.05]	0.02[0,0.04]
g(6)	0.01[0,0.06]	0.05[0.02,0.09]	0.01[0,0.04]	0.02[0,0.04]	0.01[0,0.03]
mean SI	2.6[2.2,3.5]	2.5[2.0,3.3]	2.6[2.1,3.3]	2.6[2.1,3.3]	2.6[2.1,3.2]
sd SI	1.3[0.9,2.4]	1.2[0.5,2.1]	1.2[0.8,2.0]	1.2[0.8,2.0]	1.2[0.8,2.0]
com. risk (10^{-4})	41[7.2,86]	56.7[14.6,126.6]	51.8[14.5,102.6]	52.9[12.4,103.9]	56.3[18.2,105.2]
child rel. sus.	1.96[1.05,3.78] (p=0.005*)	1.85[0.88,4.55] (p=0.007*)	2.6[1.3,5.77] (p=0.005*)	2.62[1.31,6.27] (p=0.006*)	2.77[1.37,6.26] (p=0.005*)
older adult rel. sus.	0.17[0.02,0.92] (p=0.031*)	0.17[0.01,1.04] (p=0.016*)	0.17[0.01,1.12] (p=0.027*)	0.19[0.02,1.17] (p=0.028*)	0.18[0.01,1.13] (p=0.032*)
max LL^	-296.52	-307.86	-297.63	-298.09	-297.24
DIC	611.91	625.83	608.81	610.01	608.09
p-value Exp-Obs$	0.985	0.0086	0.8197	0.7556	0.8319

^: Log-likelihood (LL).

$: from a simulation-based $\chi 2$ test comparing observed and expected distributions of the number of cases per household size.

*: derivation of those p-values is detailed in section 3.1 of the SM.

3.3 Community risk

We investigated whether age patterns in the data could be solely explained by the fact that children and adults were exposed to different community risks. We therefore considered the model where children and adults have different community risks but same susceptibility. The fit of this model was substantially lower than that of the baseline model (Table SM3: DIC

difference=11.31). The age of the household member was still a significant predictor of transmission when included in the model with different community risks for adults and children (p=0.01). No significant difference was detected between adult and children community risks (p=0.89).

Table SM3: *Comparison of the baseline model with models where the community risk is different for children and adults.*

	Baseline	**Com risk A-C without relative age susceptibility**	**Com risk A-C with relative age susceptibility**
g(2)	0.22[0.1,0.43]	0.23[0.1,0.45]	0.23[0.09,0.43]
g(3)	0.13[0.06,0.24]	0.13[0.05,0.24]	0.13[0.05,0.23]
g(4)	0.05[0.02,0.09]	0.06[0.02,0.1]	0.05[0.02,0.09]
g(5)	0.04[0.01,0.08]	0.04[0.01,0.09]	0.04[0.01,0.08]
g(6)	0.01[0,0.06]	0.02[0,0.07]	0.02[0,0.06]
mean SI	2.6[2.2,3.5]	2.6[2,4]	2.6[2.1,3.7]
sd SI	1.3[0.9,2.4]	1.3[0.7,3.3]	1.2[0.8,2.7]
com. risk (10^{-4}) - adult	41[7.2,86]	48.4[6.8,100.9]	44[11.8,90.8]
com. risk (10^{-4}) - child	41[7.2,86]	244.09[42.02,564.46]	162.65[6.08,490.48]
child rel. sus.	1.96[1.05,3.78]	-	1.77[0.84,3.58]
older adult rel. sus.	0.17[0.02,0.92]	-	0.16[0.01,0.88]
max LL^	-296.52	-303.28	-296.51
DIC	611.91	623.22	613.98
p-value Exp-Obs$	0.985	0.9647	0.9772

^: Log-likelihood (LL).

$: from a simulation-based $\chi 2$ test comparing observed and expected distributions of the number of cases per household size.

We also evaluated the robustness of estimates to the assumption that the community risk grows exponentially during follow-up. To that end, we modelled the community risk on day d as:

$$\alpha(d) = \alpha_0 \exp\{r(d - d_0)\}$$

where d_0 corresponds to the first day of symptom onset in the dataset (April 12th 2009) and r is the exponential growth rate assumed to be larger than 0.

Under the assumption that r>0, the estimated exponential growth rate is 0.04 ([0.002, 0.136]) (Table SM4). There was no significant improvement compared to the model with no exponential growth (p=0.33). Accounting for exponential growth in the community risk does not affect other estimates.

Table SM4: *Comparison of the baseline model with the model where the community risk grows exponentially.*

	No exponential growth rate (baseline)	Exponential growth rate
g(2)	0.22[0.1,0.43]	0.23[0.1,0.44]
g(3)	0.13[0.06,0.24]	0.13[0.07,0.23]
g(4)	0.05[0.02,0.09]	0.05[0.02,0.09]
g(5)	0.04[0.01,0.08]	0.04[0.01,0.08]
g(6)	0.01[0,0.06]	0.01[0,0.06]
mean SI	2.6[2.2,3.5]	2.6[2.1,3.5]
sd SI	1.3[0.9,2.4]	1.3[0.9,2.5]
α_0 (10^{-4})	41[7.2,86]	14.4[1,58]
exp growth rate r (r>0)	-	0.043 [0.002,0.136]
child rel. sus.	1.96[1.05,3.78]	1.97[1.07,3.67]
older adult rel. sus.	0.17[0.02,0.92]	0.15[0.01,0.85]
max LL^	-296.52	-296.05
DIC	611.91	608.9
p-value Exp-Obs$	0.985	0.9847

^: Log-likelihood (LL).

$: from a simulation-based $\chi 2$ test comparing observed and expected distributions of the number of cases per household size.

3.4 Serial interval

We explored models where the median serial interval depended on the age of the infector (the shape parameter *k* was assumed to be independent of the age group; see equation 2). No significant difference was found (p=0.33; Table SM5).

3.5 Statistical approaches to deal with missing dates of onset

There were 11 households with complete age and diagnostic information, but with missing day of symptom onset – we only know that the symptom date for those individuals is within 7 days (+/-) from onset in the index case. In general, statistical methods such as data augmentation can be used to account for such uncertainty. However, here, it is not possible to know which of those households should be excluded from the analysis on the basis that the index case was not the first case in the household.

Excluding all those households from the analysis may underestimate transmission rates while including all of them may overestimate them. We have therefore decided to explore those two extreme scenarios and find that the 2 approaches give very similar results.

In the baseline scenario, households with missing dates of onset were excluded from the analysis. Below, we explore the alternative extreme scenario where all those households are included in the analysis, with all missing onset dates assumed to be in the interval {day 0,…, day 7}. In this model, we also estimate the probability p_C that a household member is a co-primary case. The missing dates of symptom onset are considered as augmented data. We explore the joint posterior distribution of augmented data and transmission parameters via

MCMC sampling. In such data augmentation frameworks, likelihood ratio tests and DIC are not available for model comparison.

As can be seen in Table SM6, the main impact of including those households in the analysis is to increase the estimated community risk of transmission. The impact on other parameters is marginal.

Table SM5: *Comparison of the baseline model with the model where the serial interval depends on the age of the infector.*

	Baseline model	**SI A-C without relative age susceptibility**	**SI A-C with relative age susceptibility**
g(2)	0.22[0.1,0.43]	0.24[0.1,0.47]	0.22[0.09,0.42]
g(3)	0.13[0.06,0.24]	0.15[0.07,0.25]	0.13[0.06,0.24]
g(4)	0.05[0.02,0.09]	0.06[0.02,0.11]	0.05[0.02,0.09]
g(5)	0.04[0.01,0.08]	0.05[0.01,0.1]	0.04[0.01,0.08]
g(6)	0.01[0,0.06]	0.01[0,0.06]	0.01[0,0.05]
mean SI- adult	2.6[2.2,3.5]	2.6[2.1,4.4]	2.6[2.2,3.4]
sd SI-adult	1.3[0.9,2.4]	1.3[0.8,3.9]	1.2[0.9,2.3]
mean SI- child	2.6[2.2,3.5]	4.3[1.1,9.7]	4.2[1.1,9.7]
sd SI- child	1.3[0.9,2.4]	2.3[0.2,6.4]	2.2[0.2,5.7]
com. risk adult (10^{-4})	41[7.2,86]	51.7[9.4,104.1]	45.9[13.6,90.3]
child rel. sus.	1.96[1.05,3.78]	-	2.01[1.06,3.89]
older adult rel. sus.	0.17[0.02,0.92]	-	0.16[0.01,0.86]
max LL$^{\wedge}$	-296.52	-304.15	-296.04
DIC	611.91	624.31	612.21
p-value Exp-Obs$^{\$}$	0.985	0.9774	0.9825

$^{\wedge}$: Log-likelihood (LL).

$^{\$}$: from a simulation-based χ^2 test comparing observed and expected distributions of the number of cases per household size.

Table SM6: *Comparison of estimates when households with missing onset are excluded from the analysis (baseline scenario) and when they are included in the analysis.*

	Exclude (baseline)	**Include**
g(2)	0.22[0.1,0.43]	0.25[0.11,0.47]
g(3)	0.13[0.06,0.24]	0.13[0.06,0.23]
g(4)	0.05[0.02,0.09]	0.06[0.02,0.1]
g(5)	0.04[0.01,0.08]	0.04[0.01,0.08]
g(6)	0.01[0,0.06]	0.01[0,0.06]
mean SI	2.6[2.2,3.5]	2.6[2.2,3.3]
sd SI	1.3[0.9,2.4]	1.2[0.9,2.1]
com. risk (10^{-4})	41[7.2,86]	53.3[15.7,104.5]
child rel. sus.	1.96[1.05,3.78]	1.89[1.03,3.58]
older adult rel. sus.	0.17[0.02,0.92]	0.14[0.01,0.82]
proba coprimary case (10^{-2})	-	2.8[1.7,4.4]
p-value Exp-Obs$^{\$}$	0.985	0.9988

$^{\$}$: from a simulation-based χ^2 test comparing observed and expected distributions of the number of cases per household size.

3.6 Priors

To ensure good convergence of the MCMC chain, we used the well-documented prior information that less than *P*=20% of transmission occur after day 14 (day 0 being defined as symptom onset in the case) [3]. Table SM7 shows that estimates would be unchanged for *P*=30% or *P*=40%. If we do not use such a prior (P=100%), the MCMC chain is sometimes "trapped" in a local maximum. This area of the parameter space was characterized by i) a very small maximum log-likelihood (-321.87 as opposed to about -296.5 in Table SM7, where P≤40%); ii) implausible parameter values with the mean serial interval diverging to infinity. The prior on *P* was therefore used as a way to avoid this technical issue. An alternative option would have been to design a more efficient set of MCMC updates so that it would be possible to move from the local maximum to the real one. However, given the very low log-likelihood values associated with the local minimum, it seemed more appropriate and simple to slightly restrict the parameter space in a range of plausible values.

We also explore the impact of the prior on the relative susceptibility parameters by modifying the variance of the log-normal distribution. Results are presented in Table SM8. The prior has little impact on the posterior distribution of child relative susceptibility. As expected, it has a larger impact on the posterior distribution of the relative susceptibility of older adults since the number of older adults among household contacts is smaller. The p-values presented in section 3.1 and 3.2 are not affected by those priors.

Table SM7: *Sensitivity of estimates to the prior assumption that less than P=20% of transmission occur after day 14.*

	P=20% (baseline)	P=30%	P=40%
g(2)	0.22[0.1,0.43]	0.23[0.09,0.44]	0.23[0.09,0.45]
g(3)	0.13[0.06,0.24]	0.13[0.07,0.24]	0.14[0.07,0.25]
g(4)	0.05[0.02,0.09]	0.05[0.02,0.09]	0.05[0.02,0.09]
g(5)	0.04[0.01,0.08]	0.04[0.01,0.08]	0.04[0.01,0.09]
g(6)	0.01[0,0.06]	0.01[0,0.06]	0.02[0,0.06]
mean SI	2.6[2.2,3.5]	2.6[2.2,3.7]	2.6[2.2,4.1]
sd SI	1.3[0.9,2.4]	1.3[0.9,2.8]	1.3[0.9,3.3]
com. risk (10^{-4})	41[7.2,86]	42.2[9.5,87.2]	40.7[8.6,85]
child rel. sus.	1.96[1.05,3.78]	1.97[1.06,3.69]	1.94[1.04,3.64]
older adult rel. sus.	0.17[0.02,0.92]	0.15[0.02,0.81]	0.15[0.01,0.82]
max LL$^{\wedge}$	-296.52	-296.53	-296.49
DIC	611.91	611.77	611.74
p-value Exp-Obs$^{\$}$	0.985	0.9842	0.9848

$^{\wedge}$: Log-likelihood (LL).

$^{\$}$: from a simulation-based χ^2 test comparing observed and expected distributions of the number of cases per household size.

Table SM8: *Sensitivity analysis with respect to the prior on the age-specific susceptibility parameters.*

	LN(0,1)	LN(0,2)	LN(0,3) (baseline)	LN(0,4)	LN(0,5)
[2.5%,97.5%] percentile of prior	[0.14,7.10]	[0.06,15.99]	[0.03,29.81]	[0.02,50.40]	[0.01,80.05]
g(2)	0.23[0.1,0.44]	0.22[0.09,0.43]	0.22[0.1,0.43]	0.23[0.1,0.44]	0.23[0.1,0.43]
g(3)	0.13[0.06,0.23]	0.13[0.06,0.24]	0.13[0.06,0.24]	0.13[0.07,0.24]	0.13[0.07,0.23]
g(4)	0.05[0.02,0.09]	0.05[0.02,0.09]	0.05[0.02,0.09]	0.05[0.02,0.09]	0.05[0.02,0.09]
g(5)	0.04[0.01,0.08]	0.04[0.01,0.08]	0.04[0.01,0.08]	0.04[0.01,0.08]	0.04[0.01,0.08]
g(6)	0.01[0,0.06]	0.01[0,0.06]	0.01[0,0.06]	0.01[0,0.06]	0.02[0,0.06]
mean SI	2.7[2.2,3.7]	2.6[2.2,3.5]	2.6[2.2,3.5]	2.6[2.2,3.4]	2.6[2.2,3.6]
sd SI	1.3[0.9,2.7]	1.3[0.9,2.4]	1.3[0.9,2.4]	1.3[0.9,2.3]	1.3[0.9,2.5]
com. risk	41.6[9.2,86.5]	42.3[8.5,88.5]	41[7.2,86]	42.3[11.3,87.7]	41.7[8,88.4]
child rel. sus.	1.91[1.05,3.47]	1.96[1.04,3.85]	1.96[1.05,3.78]	1.97[1.04,3.73]	1.97[1.04,3.66]
older adult rel. sus.	0.31[0.07,1.04]	0.2[0.02,0.91]	0.17[0.02,0.92]	0.11[0.01,0.81]	0.11[0,0.79]
DIC	612.92	612.38	611.91	611.74	611.98
p-value Exp-Obs$^{\$}$	0.9841	0.9839	0.985	0.9846	0.9853

$^{\wedge}$: Log-likelihood (LL).

$^{\$}$: from a simulation-based χ^2 test comparing observed and expected distributions of the number of cases per household size.

3.7 Models for infectivity

Along with the baseline model, we explored 7 model variants to evaluate if any of the following variables were associated with significant increase in infectivity: age of the case, cough, fever, runny nose sore throat or diarrhea in the case. Table SM9 summarizes the findings. The likelihood ratio test indicates that none of the effects were significant predictors of infectivity. Besides, none of the models substantially improved the DIC.

It is possible that some variations in infectiousness are not captured by the age or the symptoms of the case. The resulting over-dispersion could affect estimates of other parameters. However, such problem would be expected to affect the final size distribution (Table 3). This does not seem to be the case since the model with constant infectivity gives a very good fit to the final size distribution (Table 3).

Table SM9: *Models assessing potential predictors for increased infectivity.*

	Baseline	**Age**	**Cough**	**Fever**	**Runny nose**	**Sore throat**	**Diarrhea**
g(2)	0.22[0.1,0.43]	0.27[0.11,0.52]	0.23[0.1,0.46]	0.22[0.1,0.44]	0.29[0.13,0.56]	0.23[0.1,0.46]	0.22[0.1,0.42]
g(3)	0.13[0.06,0.24]	0.18[0.08,0.34]	0.14[0.07,0.25]	0.13[0.06,0.24]	0.16[0.07,0.3]	0.14[0.06,0.25]	0.14[0.06,0.23]
g(4)	0.05[0.02,0.09]	0.07[0.03,0.14]	0.05[0.02,0.09]	0.05[0.02,0.09]	0.06[0.02,0.11]	0.05[0.02,0.1]	0.05[0.02,0.09]
g(5)	0.04[0.01,0.08]	0.05[0.02,0.13]	0.04[0.01,0.08]	0.04[0.01,0.08]	0.05[0.02,0.11]	0.04[0.01,0.09]	0.04[0.01,0.08]
g(6)	0.01[0,0.06]	0.02[0,0.1]	0.01[0,0.05]	0.01[0,0.06]	0.02[0,0.08]	0.01[0,0.07]	0.01[0,0.06]
mean SI	2.6[2.2,3.5]	2.6[2.1,3.4]	2.6[2.2,3.4]	2.6[2.1,3.5]	2.6[2.2,3.4]	2.6[2.2,3.4]	2.6[2.2,3.5]
sd SI	1.3[0.9,2.4]	1.2[0.9,2.3]	1.2[0.9,2.2]	1.2[0.9,2.4]	1.2[0.9,2.2]	1.3[0.9,2.3]	1.3[0.9,2.5]
com. risk (10^{-4})	41[7.2,86]	46.6[11.2,91.7]	45[12.8,89.9]	45.2[9.8,91.5]	44.1[11.7,92]	43.7[10.3,91.5]	41.7[8.3,87.5]
child rel. sus.	1.96[1.05,3.78]	1.93[1.01,3.69]	1.97[1.05,3.82]	1.98[1.05,3.83]	2.02[1.05,3.92]	1.96[1.05,3.72]	1.96[1.06,3.74]
older adult rel. sus.	0.17[0.02,0.92]	0.14[0.01,0.84]	0.15[0.01,0.82]	0.15[0.01,0.87]	0.13[0.01,0.76]	0.15[0.01,0.88]	0.16[0.01,0.88]
child rel. inf.		0.55[0.26,1.05]					
older adult rel. inf.		0.21[0.02,1.2]					
no cough vs cough			0.36[0.03,1.37]				
no fever vs fever				0.44[0.02,3.01]			
no runny nose vs runny nose					0.46[0.18,0.88]		
no sore throat vs sore throat						0.85[0.39,1.61]	
diarrhea vs no diarrhea							0.99[0.04,27.95]
max LL^	-296.52	-294.53	-295.76	-296.4	-295.06	-296.46	-296.44
DIC	611.91	611.69	611.06	612.33	611.13	614.14	611.78
p-value Exp-Obs$	0.985	0.989	0.9236	0.9529	0.9665	0.9809	0.985
p (likelihood ratio test)		0.137	0.218	0.624	0.087	0.729	0.689

^: Log-likelihood (LL).

$: from a simulation-based $\chi 2$ test comparing observed and expected distributions of the number of cases per household size.

3.8 Clinical definition of a case

Table SM10 and SM11 show how estimates would change if ILI was used rather than ARI as a clinical definition of a case.

Table SM10: *Estimates obtained when the case definition is ARI (baseline) and when it is ILI.*

	ARI (baseline)	**ILI**
g(2)	0.22[0.1,0.43]	0.14[0.05,0.29]
g(3)	0.13[0.06,0.24]	0.1[0.05,0.18]
g(4)	0.05[0.02,0.09]	0.04[0.02,0.07]
g(5)	0.04[0.01,0.08]	0.04[0.02,0.08]
g(6)	0.01[0,0.06]	0.0045[0.0001,0.03]
mean SI	2.6[2.2,3.5]	2.6[2.1,4.6]
sd SI	1.3[0.9,2.4]	1.4[1,4.5]
com. risk	41[7.2,86]	12.3[1.2,39.7]
child rel. sus.	1.96[1.05,3.78]	2.44[1.36,4.5]
older adult rel. sus.	0.17[0.02,0.92]	0.17[0.01,1.05]
max LL^	-296.52	-249.71
DIC	611.91	518.07
p-value Exp-Obs$	0.985	0.9993

^: Log-likelihood (LL).

$: from a simulation-based $\chi 2$ test comparing observed and expected distributions of the number of cases per household size.

Table SM11: *Odds ratio estimates (and 95% confidence intervals) for onset of ILI in household contacts. These results come from a logistic GEE model including age group of the household member and household size (log-transformed).*

		Odds ratio estimate	**95% confidence interval**		**p**
			Lower bound	**Upper bound**	
Age of household member	0-4 y	6.20	2.59	14.83	<0.001
	5-18 y	3.00	1.52	5.93	0.002
	19-50 y	1			
	≥51 y	0.38	0.05	2.91	0.35
Doubling of household size		0.23	0.11	0.49	<0.001

3.9 Dataset

Table SM12 and SM13 are the same as Table 2 and Table 3 of the main article but for the 339 households of size 2-6 in which the index case was the first case of the household, with no missing information on symptoms but in which information on age may be missing.

Table SM12: *Demographic and clinical characteristics of the household contacts in 339 households of size 2-6 in which the index case was the first case of the household and there was no missing information on symptoms.*

	Total	0-23 mo	2-4 yr	5-18 yr	19-50 yr	≥51 yr
Number	968	17 (2%*)	37 (5%*)	237 (34%*)	357 (51%*)	54 (8%*)
Median Age	21					
Clinical symptoms						
Fever or feverish	98 (10%[§])	2 (12%[§])	12 (32%[§])	39 (16%[§])	35 (10%[§])	2 (4%[§])
Cough	127 (13%[§])	3 (18%[§])	11 (30%[§])	42 (18%[§])	53 (15%[§])	1 (2%[§])
Sore throat	62 (6%[§])	2 (12%[§])	3 (8%[§])	19 (8%[§])	31 (9%[§])	2 (4%[§])
Runny nose	57 (6%[§])	3 (18%[§])	7 (19%[§])	20 (8%[§])	18 (5%[§])	2 (4%[§])
Diarrhea	19 (2%[§])	2 (12%[§])	1 (3%[§])	9 (4%[§])	7 (2%[§])	0 (0%[§])
ILI[♦]- no (%)	75 (8%[§])	2 (12%[§])	10 (27%[§])	31 (13%[§])	26 (7%[§])	1 (2%[§])
RR[$] [95% CI] (p)			4.47[2.45,8.16] (p<0.001)	1.80[1.09,2.96] (p=0.02)	1	0.326[0.049,2.153] (p=0.24)
ARI^- no (%)	100 (10%[§])	2 (12%[§])	12 (32%[§])	37 (16%[§])	39(11%[§])	2 (4%[§])
RR[$] [95% CI] (p)		1.07[0.26,4.45] (p=0.92)	2.96 [1.83,4.77] (p<0.001)	1.51 [1.06,2.16] (p=0.023)	1	0.32 [0.078,1.32] (p=0.11)

[♦]: Influenza-Like Illness (ILI) defined as fever/feverishness and (cough and/or sore throat).

^: Acute Respiratory Illness (ARI) defined as at least two of the following signs: fever/feverishness, cough, sore throat, runny nose.

[$]: Relative Risk (RR).

Percentages are relative number of index cases or household contacts within that column, except for the percentages reported in the rows labeled "Male sex", "Hospitalized" and "Clinical symptoms" (denoted %[§]) in which the percentages relate to the number of index cases of household contacts within that column with the relevant information and the percentages reported in the rows labeled "Number" (denoted %*) in which the percentages relate to the total number of index cases or household contacts with age information.

Table SM13: *Odds ratio estimates (and 95% confidence intervals) for onset of ARI in household members (excluding those with ARI onset before the index case onset). These results come from a logistic GEE model including age group of the household member and household size (log-transformed).*

		Odds ratio estimate	95% confidence interval		p
			Lower bound	Upper bound	
Age of household member	0-4 y	3.97	1.93	7.74	<0.001
	5-18 y	1.88	1.12	3.18	0.02
	19-50 y	1			
	≥51 y	0.33	0.06	1.66	0.18
Doubling of household size		0.28	0.15	0.52	<0.001

4 Laboratory confirmation of index cases

Confirmed and probable cases were tested using CDC developed assays. The CDC developed a RT-PCR assay to detect seasonal influenza A, B, H1, H3, and avian H5 serotypes. This assay was approved by the Food and Drug Administration (FDA) and was distributed to public health laboratories in December 2008. Primers and probes specific for swine influenza A (H1) were recently developed and tested for use in a modified version of this assay for the detection of swine influenza A (H1 and H3) subtypes. In order to develop a test specific for the detection of novel influenza A(H1N1) virus infection, CDC modified reagents previously developed for the swine influenza A (H1 and H3) assay. Novel influenza A(H1N1) RT-PCR kits were distributed to public health laboratories in early May.

References:

1. Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boelle PY. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Stat Med. 2004 Nov 30;23(22):3469-87.
2. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman and Hall; 1996.
3. Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, Leach S, et al. Time lines of infection and disease in human influenza: A review of volunteer challenge studies. Am J Epidemiol. 2008 Apr 1;167(7):775-85.
4. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Statist Soc B. 2002;64(4):583--639.