Influenza transmission in households during the 1918 pandemic

Christophe Fraser[1], *Derek A. T. Cummings*[2], *Don Klinkenberg*[1,3],

Donald S. Burke[4] *and Neil M. Ferguson*[1]

[1]Medical Research Council Centre for Outbreak Modelling and Analysis, Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, London W2 1PG, UK, [2]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205,USA, [3]Theoretical Epidemiology, Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands, [4]Graduate School of Public Health, University of Pittsburgh, USA.

Web Appendix

Consists of copies of the historical documents used in this study, detailed descriptions of the methods, and supplementary results and sensitivity analyses

Locality 1.
Baltimore

no. of households of each stated size having stated no. of cases.

Balto.

Cases in House-hold	No. of persons in Household															Total House-holds	Total People	Total Cases
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	17			
0	207	807	767	703	436	288	141	113	45	22	10	10	2	2		3553	13692	0.
1	37	228	351	296	218	136	91	32	18	11	3	2	1			1424	5591	1424
2		91	140	211	161	84	47	35	20	6	1	1			1	798	3644	1596
3			84	91	81	76	42	24	9	6	1		1			415	2128	1245
4				77	65	41	32	27	5	1	3	2				253	1431	1012
5					56	28	36	13	8	7	3.					151	979	755
6						27	21	13	5	2		2				70	507	420
7							24	12	3	9		1.				49	393	343
8								15	4	6			1			26	229	209
9										2	2	2				6	66	54
10										2	2	1		1		6	68	66
11												1				1	12	11
12												1				1	12	12
Total H. H.	244	1126	1342	1378	1017	680	434	284	117	74	25	23	5	3	1	6753	28977	7140
Total People	244	2252	4026	5512	5085	4080	3038	2272	1053	740	275	276	65	42	17		28977	
Total Cases	37	410	883	1299	1323	998	913	629	228	241	73	82	12	10	2			
%	15.3	18.4	21.9	22.0	25.9	24.4	30.0	27.6	21.6	32.5	26.5	29.7	18.4	23.8	11.8			7140

LOCALITY - 2

FREDERICK.

NUMBER OF HOUSEHOLDS OF EACH STATED SIZE HAVING STATED NUMBER OF CASES.

Cases in Household	Number of Persons in Household:														
	1	2	3	4	5	6	7	8	9	10	11	12	Total Households	Total People	Total Cases
0	17	58	67	60	35	15	10	5	1	1	1	1	271	991	0
1	3	26	34	17	7	3	6			1			97	330	97
2		14	16	16	7	6	4	1					64	247	128
3			7	6	7	4	1	1	1			1	28	140	84
4				5	15	4	4	3		2			33	191	132
5					10	2	3	1	1	1		1	19	122	95
6						2	2			1			5	36	30
7							3	3	1		1		8	65	56
8								1	4	1	2		8	76	64
9									1				1	9	9
Total Household	20	98	124	104	81	36	33	15	9	7	4	3	534	2207	695

INFLUENZA SUBSEQUENT ATTACK RATES IN CANVASSED HOUSEHOLDS [illegible]
IN WHICH THERE WERE CASES, BY AGE GROUP AND SIZE OF HOUSEHOLD.1918.

Table found the
(c) Cases in report

Number of persons in household	Age group						
	Total	-5	5 - 9	10-19	20-29	30-39	40+
	Population of attacked households (less primary cases)						
All households	11,020	1,350	1,146	2,036	1,779	1,737	2,972
1	-	-	-	-	-	-	-
2	291	1	-	9	78	69	134
3	1,062	148	52	84	236	230	312
4	1,852	219	141	199	331	327	615
5	2,072	271	191	319	336	377	578
6	1,802	221	219	403	252	270	437
7	1,616	204	223	382	219	198	390
8	1,069	125	161	298	137	121	227
9	516	57	66	164	67	44	118
10	425	64	55	107	75	53	71
11	145	19	21	31	19	21	34
12	132	21	13	25	21	17	35
13	30	-	-	9	4	4	13
14	13	-	2	5	2	1	3
17	15	-	2	1	2	5	5
	Number of cases subsequent to the primary case						
All households	2,825	450	451	581	460	451	412
1	-	-	-	-	-	-	-
2	63	1	-	3	25	12	22
3	220	33	20	16	59	58	34
4	431	69	47	62	103	84	66
5	490	92	76	80	89	87	66
6	448	79	73	99	56	67	74
7	478	73	99	115	68	62	61
8	330	44	74	93	38	41	40
9	96	11	16	32	9	12	16
10	146	29	24	48	17	13	15
11	53	5	11	15	6	9	7
12	58	14	9	11	7	6	11
13	3	-	-	2	1	-	-
14	9	-	2	5	2	-	-
17	-	-	-	-	-	-	-
	Subsequent attack rate per 100						
All households	25.7	33.3	39.3	28.5	27.0	26.0	13.9
1	-	-	-	-	-	-	-
2	21.6	100.0	-	33.3	32.0	17.4	16.4
3	20.7	22.3	38.5	19.0	25.0	25.2	10.9
4	23.0	31.5	33.1	31.2	31.1	25.7	10.7
5	23.7	33.9	39.8	25.1	26.5	23.1	11.4
6	24.9	35.8	33.3	24.6	22.2	24.8	16.9
7	29.6	35.8	44.4	30.0	31.0	31.3	15.6
8	30.9	35.2	46.0	31.2	27.7	33.9	17.6
9	18.6	19.3	24.2	19.5	13.4	27.3	13.6
10	34.3	45.3	43.6	44.9	22.7	24.5	21.1
11	36.5	26.3	52.4	48.4	31.6	42.9	20.6
12	44.0	66.7	69.2	44.0	33.3	35.3	31.4
13	10.0	-	-	22.2	25.0	-	-
14	69.2	-	100.0	100.0	100.0	-	-
17	-	-	-	-	-	-	-

Description of the study

The background historical context to the Baltimore-based Frost-Sydenstricker study was described in Chapter 7 of (1), and the first detailed description of the study was given in (2) with some further analyses focusing on the impact of socio-economic status in (3).

Wade Hampton Frost worked first as a field epidemiologist in the United States Public Health Service and later, in 1919, was the founding chair of the department of epidemiology at Johns Hopkins School of Public Health (1). In 1917, the Public Health Service was incorporated into the US military, and Frost noted that *"the conditions of war…impose additional burdens…as a result of the concentration of population in and around military encampments and industrial centers. Additional public health problems will arise in the civil population as the war progresses…"* (1) which seems prescient given accounts of how the H1N1 pandemic emerged (4). In April 1918, during what is now recognized as the herald spring wave of the pandemic epidemic, an Office of Field Investigations of Influenza was organized and led by Frost (1). The task of this office was to collate and analyze influenza statistics. Following the far more devastating autumn wave of the epidemic, the present study of influenza transmission in households in Maryland (2) appears to have been a pilot study to establish influenza infection rates which could not be deduced from weekly mortality data (1). The study design, based on systematic household canvassing, was pioneered by Sydenstricker who in 1919 became the chief statistician of the US Public Health Service (1).

After describing the localities in Maryland where surveys were carried out (2), the study design was described as follows:

"*In each of these localities a numbers of areas were selected for house-to-house canvass. The size of the areas was roughly fixed so as to include approximately the same number of persons in each, the selection of the areas within a town or city being made so as to give a fairly uniform geographic distribution. Enumerators were employed to visit every house in the selected areas, and to make enquiries of the housewife or other responsible member of the household as to the sex and age of each member, the date of onset and duration of each case of influenza or pneumonia, and the date of each death from influenza or pneumonia.*

"Thus the sex and age of every person in the population canvassed, as well as persons affected by the disease, were ascertained. The data obviously are crude to a certain degree because of the following conditions: (1) The canvass was made some time after the earlier cases occurred, and the dates of onset were not accurately recalled for a small proportion of the cases; (2) the families' statements as to diagnosis were accepted; (3) the enumerators were not persons especially trained in this work, although they were carefully selected for intelligence and reliability. […]"

Of relevance here, the canvass in Baltimore was carried out Nov. 20 to Dec. 15 1918, and in Frederick in Nov. 27 to 30. In 1917, as reported in (2), the total population of Baltimore was 594,637 and that of Frederick 11,225. This is in close agreement with the census population size of 599,653 inhabitants for Baltimore reported in (5). In total 33,776 people were reported to have been included in the Baltimore canvass and 2,420 in the Frederick canvass (2).

Approximately 3,927 (0.65%) of the city's population of 599,653 inhabitants died of pneumonia or influenza in eight weeks (5, 6).

For the current study, new documents pertaining to the Maryland canvasses were identified in the Frost collection of the Chesney Medical Archives at the Johns Hopkins University (Document S1). These identify data from 6,774 households in Baltimore and 534 households in Frederick.

S1. The data

From the data in Document S1, we can tabulate the number of households of size n reporting m cases, which we denote $k_{(m,n)}$, in Tables S1 and S2.

		Number of persons in household, *n*														
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**17**
Cases in household, *m*	**0**	207	807	767	703	436	288	141	113	45	22	10	10	2	2	0
	1	37	228	351	296	218	136	91	32	18	11	3	2	1	0	0
	2		91	140	211	161	84	47	35	20	6	1	1	0	0	1
	3			84	91	81	76	42	24	9	6	1	0	1	0	0
	4				77	65	41	32	27	5	1	3	2	0	0	0
	5					56	28	36	13	8	7	3	0	0	0	0
	6						27	21	13	5	2	0	2	0	0	0
	7							24	12	3	9	0	1	0	0	0
	8								15	4	6	0	0	1	0	0
	9									0	2	2	2	0	0	0
	10										2	2	1	0	1	0
	11											0	1	0	0	0
	12												1	0	0	0

Table S1 - the distribution households according to size and number of influenza cases, based on the canvass in Baltimore

		Number of persons in household, *n*											
Cases in household, *m*		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
	0	17	58	67	60	35	15	10	5	1	1	1	1
	1	3	26	34	17	7	3	6	0	0	1	0	0
	2		14	16	16	7	6	4	1	0	0	0	0
	3			7	6	7	4	1	1	1	0	0	1
	4				5	15	4	4	3	0	2	0	0
	5					10	2	3	1	1	1	0	1
	6						2	2	0	0	1	0	0
	7							3	3	1	0	1	0
	8								1	4	1	2	0
	9									1	0	0	0
	10										0	0	0
	11											0	0
	12												0

Table S2 – the distribution households according to size and number of influenza cases, based on the canvass in Frederick

The data from Baltimore are plotted in Fig. 1B, along with some summary statistics. Here (Fig S1) we show the data plotted for both Baltimore and Frederick for direct comparison.

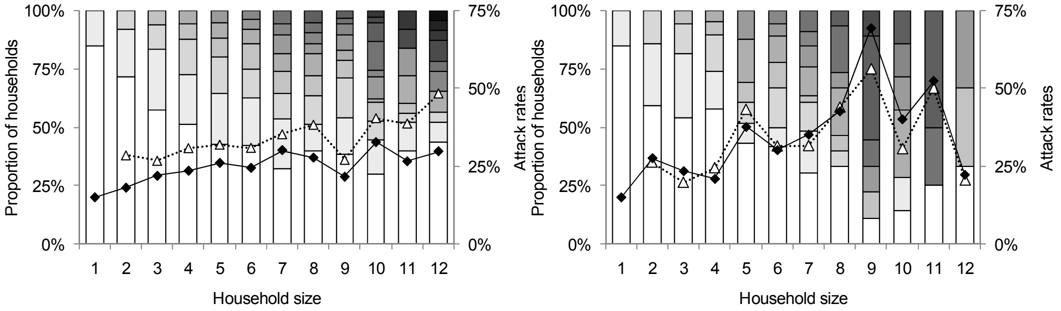

Figure S1 – For each household size, the bars show the proportion of households recording 0, 1, 2, etc... cases, highlighted in darkening shades of gray. a, is identical to Fig. 1b, while b shows the data for Frederick. The black symbols show the mean attack rate for each household size, while the open symbols show the secondary attack rate. Data from households of size >12 are sparse and are not shown.

In Figure S1, the mean attack rate for each household size is calculated as $AR(n) = \left(\sum_{m=0}^{n} m\, k_{(m,n)}\right) \Big/ \left(n \sum_{m=0}^{n} k_{(m,n)}\right)$, while the secondary attack rate is calculated as $SAR(n) = \left(\sum_{m=1}^{n} (m-1) k_{(m,n)}\right) \Big/ \left((n-1) \sum_{m=1}^{n} k_{(m,n)}\right)$.

The largest household sampled was of size $n_{\max} = 17$. The number of households sampled was $N_H = \sum_{n=1}^{n_{\max}} \sum_{m=0}^{n} k_{(m,n)}$, while the number of people sampled was $N = \sum_{n=1}^{n_{\max}} \sum_{m=0}^{n} n\, k_{(m,n)}$. The total number of cases reported was $M = \sum_{n=1}^{n_{\max}} \sum_{m=0}^{n} m\, k_{(m,n)}$. For Baltimore: $N_H = 6,753$, $N = 28,977$ and $M = 7,140$; for Frederick: $N_H = 534$, $N = 2,207$ and $M = 695$.

The distribution of household sizes is denoted $pr[n] = (1 / N_H) \sum_{m=0}^{n} k_{(m,n)}$.

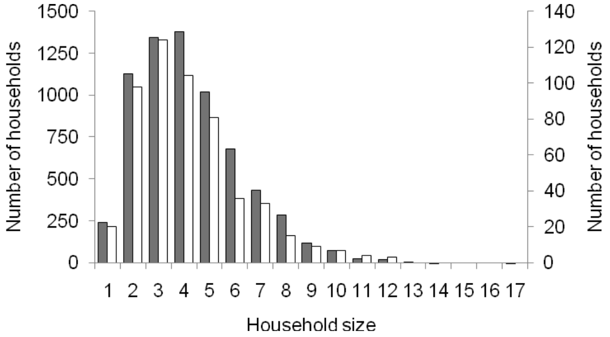

Figure S2 – the distribution of households plotted by size for Baltimore (dark bars, left axis) and Frederick (light bars, right axis).

Based on this distribution, the mean household size is $E[n] = \sum_{n=1}^{n_{\max}} n\, pr[n] = N / N_H$. In Baltimore, the mean household size is $E[n] = 4.29$ while in Frederick it is $E[n] = 4.13$. The probability that an individual randomly chosen from the population lives in a household of size n, denoted $f[n]$, is given by the size-biased household size distribution, $f[n] = n\, \mathrm{pr}[n] / E[n]$.

The estimated overall attack rate, $\mathrm{AR} = M / N$, is $\mathrm{AR} = 24.6\%$ (95% confidence interval (c.i.) [24.2%-25.2%]) for Baltimore and $\mathrm{AR} = 31.5\%$ (95% c.i. [29.6%- 33.5%]) for Frederick.

The original report of the study (2) gave $N = 33,776$, $M = 7,868$ for Baltimore and $N = 2,420$, $M = 777$ for Frederick (from Table 1 in (2)). The data reported in Document S1 thus does not include all the households in the original report. The size of the data in Document S1 represent 85.8%, 90.7%, 91.2% and 89.4% of the original published study's sample sizes, respectively; the inclusion criteria from the original published reports are unclear. The attack rates in the original report were similar to those derived from the present data, i.e. $\mathrm{AR} = 23.3\%$ for Baltimore and $\mathrm{AR} = 32.1\%$ for Frederick estimated in (2).

S2. The Reed-Frost transmission model

The Reed-Frost model is formulated in discrete generations of infection (7). The model describes the spread of an infection which generates sterilizing immunity within a closed population, and is defined by two rules. The first is that no one is re-infected. The second is that, if susceptible, there is a constant probability of being infected by each infectious individual in the previous generation. This probability, denoted p, is sometimes known as the susceptible-infectious transmission probability, and $q = 1 - p$ is known as the escape probability. Confusingly, p is sometimes also known as the secondary attack rate; here we reserve secondary attack rate for the total proportion of the household which is infected following the introduction of an infectious case.

These rules lead to a chain binomial model of the number of new infections in each generation. Let t denote the generation of infection $\left(t = 0,1,2\ldots\right)$, and s_t and i_t denote the number of susceptible and infectious individuals, respectively. The probability that there are i_{t+1} infectious individuals in the next generation is then

$$\mathrm{pr}\left(i_{t+1}\right) = \begin{pmatrix} s_t \\ i_{t+1} \end{pmatrix}\left[1 - q^{i_t}\right]^{i_{t+1}}\left[q^{i_t}\right]^{s_t - i_{t+1}} \qquad [1]$$

and the number of susceptible individuals is reduced to $s_{t+1} = s_t - i_{t+1}$. For small population sizes, as in households, this model can be solved to give closed equations for distribution $P_m^{\left(s_0,i_0\right)}$ of the final number infected, $m = s_0 - s_\infty$, with s_0 and i_0 initial susceptibles and infecteds, respectively (8).

The usefulness of this simple discrete generation model is that it can represent far more complex infection processes due to a fundamental superposition principal for epidemics, which states that the order in which 'infectious exposures' take place in an epidemic does not affect the final number of individuals infected (8, 9). An 'infectious exposure' refers to an event which leads to infection if and only if the recipient of the exposure is susceptible. The discrete-generation Reed-Frost model can thus be viewed as a convenient mathematical construct used to compute the final size of more complex real-time outbreaks.

More precisely, given a fixed probability of infectious exposure from outside the household $\left(1 - Q\right)$ for each household member, and a particular distribution of within-household infectiousness, the distribution of final sizes can be correctly calculated on the assumption that all outside potentially infectious exposures happen simultaneously and before any within-household transmission, even if in fact they occur at different times and possibly after there has been within-household transmission.

The initial number infected i_0 can thus be chosen to represent the number of individuals infected outside of the household, at any stage in the epidemic. Given the definition of Q above, the probability distribution of initial values i_0 in a household of size n is given by the binomial distribution

$$pr\left(i_0\right) = \binom{n}{i_0}\left[1 - Q\right]^{i_0}\left[Q\right]^{n-i_0} \qquad [2]$$

and the initial number of susceptibles is $s_0 = n - i_0$. Our model is not 'closed', in the sense that we do not link the escape probability Q to the number of people infected, as in (10).

The predicted probability distribution F_m^{n,s_0} for the final number of cases, m, in a household of size n with s_0 initial susceptible individuals can be obtained as a function of the parameters Q and q which quantify the intensity of between and within household transmission, respectively[7,9-11]. Several analytical approaches exist for determining the final size distribution F_m^{n,s_0} for this model, but we use the easily solved upper triangular system of equations

$$\binom{s_0}{k} = \sum_{m=0}^{k}\binom{s_0 - m}{k - m}F_m^{n,s_0}\Big/\left\{q^{m\left(s_0-k\right)}Q^{s_0-k}\right\} \quad \text{for} \quad k = 0,\ldots,s_0 \qquad [3]$$

S3. Some generalizations of the Reed-Frost model

The model was extended to incorporate additional features of influenza transmission which we wanted to test against these data.

3.1. Heterogeneous infectiousness

The first generalization of the basic Reed-Frost model which we considered was to allow for intrinsic variability in individuals' overall infectiousness, reflecting differences in viral shedding, duration of infectiousness, and contact rates. A convenient formalism is to specify a distribution for the cumulative infection hazard to each other member of the household, denoted $h_n > 0$, possibly dependent on the size n of the household. The escape probability from an infected individual is $q\left(h_n\right) = \exp\left(-h_n\right)$, and the susceptible-infectious transmission probability is $\text{SITP}_n = 1 - q\left(h_n\right)$. The probability distribution of h is denoted $pr\left(h\right)$. Equation [3] generalizes to

$$\binom{s_0}{k} = \sum_{m=0}^{k}\binom{s_0 - m}{k - m}F_m^{n,s_0}\left[Q\right]\Big/\left\{\phi_n\left(s_0 - k\right)^m Q^{s_0-k}\right\} \quad \text{for} \quad k = 0,\ldots,s_0 \qquad [4]$$

where $\phi_n\left(x\right) = E\left[\exp\left(-h_n x\right)\right]$ is the moment generating function of the distribution of hazards in households of size n (8, 10-12) and we have now made the final size distribution $F_m^{n,s_0}\left[Q\right]$ an explicit function of Q for reasons apparent below. Note that the susceptible-infectious transmission probability is related to the moment generating function by $\text{SITP}_n = 1 - \phi_n\left(1\right)$.

To ensure a relatively general range of possibilities, we assume that the cumulative hazard h_n is distributed as a Gamma distribution with mean B_n and shape parameter k, i.e.

$$pr\left(h\right) = \frac{h^{k-1} \exp\left(-hk \;/\; B_n\right)}{\left(B_n \;/\; k\right)^k \Gamma\left(k\right)} \qquad [5]$$

This has moment generating function

$$\phi_n\left(x\right) = \left(\frac{k}{k + xB_n}\right)^k \qquad [6]$$

The function B_n may be equal to a constant parameter, i.e. $B_n = \beta$, or may be specified as a function of household size, as below.

3.2. Household size dependence of infectiousness

Previous studies have indicated that the infection hazard for influenza may be dependent on the size of the household (13, 14). To reflect this possibility, we allowed for a decreasing function of the form $B_n = \beta \big/ n^\alpha$. α is a continuous coefficient which measures how steeply the transmission hazard decreases as a function of household size (if $\alpha > 0$); $\alpha = 0$ corresponds to density-dependent transmission, while $\alpha = 1$ corresponds to frequency-dependent transmission.

Given these two model extensions, the standard Reed-Frost model is recovered when $\alpha = 0$ and $k \to +\infty$, so that the transmission hazard h is constant and independent of household size, i.e. $h = \beta$. In this case, the moment generation function becomes

$$\phi_{\text{R-F}}\left(x\right) = q^x \qquad [7]$$

where $q = \exp\left(-\beta\right)$, in which case equation [4] reduces to [3].

3.3. Prior immunity

The epidemic of H1N1 influenza virus in the autumn of 1918 may have been preceded by a spring wave of transmission with a similar virus with lower pathogenicity which could have generated immunity in part of the population (15, 16). It is also possible that resistance to infection could be generated by cross-specific immunity to other influenza (17) or non-influenza viruses.

To model this, we considered inclusion of an earlier unobserved epidemic, or series of epidemics, which generate prior immunity in the population. For parsimony, this was modelled as being identical to the main studied autumn wave in terms of within household transmission, but with a separate degree of outside exposure to reflect the extent of prior immunity. Thus, the number of individuals with prior immunity, denoted l, in households of size n was distributed according to $F_l^{n,n}\left[Q_{\text{prior}}\right]$ (given by equation [4]). Allowing for prior immunity results in a distribution of final cases in the autumn wave, denoted R_m^{n,s_0} given by

$$R_m^{n,s_0} = \sum_{l=0}^{s_0-m} F_m^{n,s_0-l}\left[Q\right]F_l^{n,s_0}\left[Q_{\text{prior}}\right] \qquad [8]$$

3.4. 'Protected' or asymptomatic uninfectious individuals

The next extension of the model allowed for a proportion of individuals p_{pr} to be either protected against infection by a mechanism not related to previous influenza transmission, or if infected, to remain un-infectious and asymptomatic. Such individuals were deemed removed from transmission in both the spring and autumn waves of transmission. Allowing for this, the distribution of cases becomes

$$S_m^{n,s_0} = \sum_{r=0}^{s_0-m} \text{Bin}\left(r, p_{\text{pr}}, n\right) R_m^{n,s_0-r} \qquad [9]$$

Where $\text{Bin}\left(i,p,n\right) = \binom{n}{i} p^i \left(1-p\right)^{n-i}$ is the standard Binomial probability distribution.

3.5. Asymptomatic infectious individuals

Next, we allowed for proportion p_{asx} of infected infectious individuals to be asymptomatic (at least to the extent of not being recorded in the Frost survey). The distribution of recorded cases is then modified to

$$T_m^{n,s_0} = \sum_{t=0}^{s_0-m} \text{Bin}\left(t, p_{\text{asx}}, m+t\right) S_{m+t}^{n,s_0} \qquad [10]$$

Note that this model can also be used to represent <100% reporting of cases, as asymptomatic infectious cases could also be thought of as symptomatic cases which were not reported.

3.6. Non-complying households

Finally, we allowed for a proportion $1 - p_{\text{com}}$ of households to either be non-compliers (in the sense of reporting nil cases for the whole household irrespective of outcome) or to be fully removed from the epidemic (due for example to effective quarantine measures, social distance from the epidemic or heritable or shared features of individuals in households). The final distribution of reported cases is then given by

$$\begin{aligned} U_0^{n,s_0} &= 1 - p_{\text{com}} + p_{\text{com}} T_0^{n,s_0} \\ U_m^{n,s_0} &= p_{\text{com}} T_m^{n,s_0} \qquad \text{for} \quad m > 1 \end{aligned} \qquad [11]$$

3.7. The full model

The full model was thus characterized by 8 parameters: the outside escape probability Q, the within-household infection hazard parameters β, α and k, the escape probability from a virus generating prior immunity Q_{prior}, the proportion asymptomatic uninfectious/protected p_{pr}, the proportion asymptomatic infectious p_{asx} and the proportion of complying households p_{com}, and from these predicted the final distribution of observed cases in households given by $P_m^n = U_m^{n,n}$.

We constructed 2^6=64 variants of the full model, representing all possible combinations of the extensions to the basic model considered. To name each model, we use a code to denominate each assumption included. All the models contain the parameters Q and β needed to define the Reed-Frost model.

- **V** denotes that k is finite, whereas by default $k \to +\infty$ (corresponding to moment generating function [7]).
- **P** denotes that $\alpha \geq 0$, whereas by default $\alpha = 0$.
- **S** denotes that $Q_{\text{prior}} \leq 1$, whereas by default $Q_{\text{prior}} = 1$.
- **X** denotes that $p_{\text{pr}} \geq 0$, whereas by default $p_{\text{pr}} = 0$.
- **A** denotes that $p_{\text{asx}} \geq 0$, whereas by default $p_{\text{asx}} = 0$.
- **R** denotes that $p_{\text{com}} \leq 1$, whereas by default $p_{\text{com}} = 1$.

So for example the model variant **PXR** had $k \to +\infty$ (fixed), $\alpha \geq 0$ (variable), $Q_{\text{prior}} = 1$ (fixed), $p_{\text{pr}} \geq 0$ (variable), $p_{\text{asx}} = 0$ (fixed) and $p_{\text{com}} \leq 1$ (variable). All models had the basic Reed Frost parameters $Q \leq 1$ (variable) and $\beta \geq 0$ (variable).

S4. Method for fitting the models

The data are a contingency table of outcomes, i.e. the number of households of size n reporting m cases, denoted $k_{(m,n)}$. We define Ω as the set of integers $\{m,n\}$ such that the contingency table is non-zero, i.e. such that $k_{(m,n)} > 0$.

Each outcome can be viewed as an independent realization of a transmission experiment, and thus the likelihood describing the goodness of fit of the model to these data is

$$L = \prod_{\{m,n\} \in \Omega} \left(P_m^n\right)^{k_{(m,n)}} \qquad [12]$$

To obtain a more objective measure of the goodness of fit, we compared this with the saturated likelihood, given by

$$L_{\text{sat}} = \prod_{\{m,n\}\in\Omega} \left(O_m^n\right)^{k_{(m,n)}} \qquad [13]$$

where $O_m^n = k_{(m,n)} \,/\, \sum_{j=0}^{n} k_{(j,n)}$ is the observed distribution of final outbreak sizes, and defined the corresponding deviance

$$\text{Dev} = 2\sum_{\{m,n\}\in\Omega} k_{(m,n)}\left(\ln\left(O_m^n\right) - \ln\left(P_m^n\right)\right) \qquad [14]$$

The model was fit by minimizing Dev with respect to all the free parameters, which is equivalent to maximizing the likelihood L. This was repeated for each of the 64 possible models. The best possible fit any model could achieve would result in $\text{Dev} = 0$.

The number of parameters which were varied to fit the model to the data ($\#\,\text{params}$) changed from model to model, and ranged from 2 for the basic Reed Frost model to 8 for the full model denoted **PVRAXS**. The number of degrees of freedom in the data ($\#\,\text{dof}$) was estimated as the number of non-zero entries in the contingency table $k_{(m,n)}$, and was 89 for the Baltimore sample and 60 for the Frederick sample.

The model equations were solved and the optimization performed using Mathematica (18).By default, the deviance was minimized using the 'FindMaximum' function with plausible initial guesses, which was quick and efficient for these models. To check for the robustness of the optimization, a selection of model fits were repeated using the 'NMaximize' function with default options and widest possible parameter ranges. In no cases did this improve the model fit, indicating that the optimization was robust and thus that the likelihood must have a relatively simple dependence on the parameters.

The only cases in which either some dependence on the initial parameters when using 'FindMaximum' and/or some difficulty in using 'NMaximize' was detected were for models containing both prior immunity (**S**) and misreporting (**R**), indicating a degree of trade-off between the corresponding parameters in the likelihood.

All the characteristics and parameters of the best fit models are described in Table S3 (below).

S5. Method for model comparison

Before performing a more rigorous model comparison exercise, the models were classified into three broad groups based on some simple criteria. Models containing the assumption of misreporting or lack of compliance with the study (assumption **R**) were treated separately, as these model variants were viewed primarily as a sensitivity analysis to understand the robustness of the model to problems with the original data.

It was also readily apparent that there was very little statistical support for the presence of asymptomatic infection (whether infectious (**A**) or uninfectious (**X**)). Therefore models containing these assumptions were grouped separately from the baseline set of models.

The different models were compared by computing the adjusted Akaike's information criterion(19) (denoted AIC_c), defined as

$$\mathrm{AIC}_c = \min\left(\mathrm{Dev}\right) + 2\left(\#\,\mathrm{params}\right)\left(\frac{\#\,\mathrm{dof}}{\#\,\mathrm{dof} - \#\,\mathrm{params} - 1}\right) + \mathrm{constant} \qquad [15]$$

where $\#\,\mathrm{parameters}$ is the number of parameters varied to minimise Dev, $\min\left(\mathrm{Dev}\right)$ is the value of Dev obtained for the best fit model, and the constant is irrelevant ($= -2\ln\left(L_{sat}\right)$, the same for each model).

Of the basic models, the best fitting model (with the lowest AIC_c, defined as AIC_c^{ref}), was model **PVS**, taken as our reference best fit model. A measure of the quality of fit relative to this reference was given by

$$\Delta\mathrm{AIC}_c = \mathrm{AIC}_c - \mathrm{AIC}_c^{ref} \qquad [16]$$

The degree of support for this model over any of its simpler variants was very strong ($\Delta\mathrm{AIC}_c > 22$, see Table S3). None of the models with asymptomatic infection fitted better than this (all had $\Delta\mathrm{AIC}_c > 0$) indicating a lack of support for these models.

Models which included misreporting (**R**) could fit better than the reference model ($\Delta\mathrm{AIC}_c < 0$), though there was still strong support for the hypotheses represented by **P**, **V** and **S**.

S6. Alternative measures of goodness of fit

To increase our understanding of the quality of fit of the different models, and their different explanatory power, we computed two alternative deviance measures based on summary statistics of the data. Note that these do not contain extra information relative to the deviance used to fit the model, but are used to explain which aspects of the data are best fit by different model variants. The first deviance was based on the proportion of households which experienced at least one case as a function of household size (the household attack rate, HAR_n), given by

$$HAR_n = \frac{\sum_{m=1}^{n} k_{(m,n)}}{\sum_{m=0}^{n} k_{(m,n)}} \qquad [17]$$

For models without prior immunity or misreporting, the predicted household attack rate is given by

$$\widehat{\mathrm{HAR}}_n = 1 - \left(1 - Q\right)^n \qquad [18]$$

which is a monotonically increasing function of household size. If misreporting is present, this becomes modified to

$$\widehat{\mathrm{HAR}}_n = p_{\mathrm{com}} \left(1 - \left(1 - Q\right)^n\right) \qquad [19]$$

which tends asymptotically to $p_{\mathrm{com}} < 1$. If prior immunity is present, then this becomes the more complex expression

$$\widehat{\mathrm{HAR}}_n = \sum_{m=0}^{n} F_m^{n,n}\left[Q_{\mathrm{prior}}\right]\left(1 - \left(1 - Q\right)^m\right) \qquad [20]$$

where $F_m^{n,s}\left[Q\right]$ is defined in equation [4] and which need not increase monotonically as a function of household size. The agreement between data and prediction is assessed by the deviance defined in analogy with [14] to be

$$\mathrm{Dev}_{\mathrm{HAR}} = 2\sum_n \left[k_{(0,n)}\left(\ln\left(1 - \mathrm{HAR}_n\right) - \ln\left(1 - \widehat{\mathrm{HAR}}_n\right)\right) + \left(\sum_{m=1}^{n} k_{(m,n)}\right)\left(\ln\left(\mathrm{HAR}_n\right) - \ln\left(\widehat{\mathrm{HAR}}_n\right)\right)\right] [21]$$

The second deviance measure was defined based on the number of cases conditional on at least one case having occurred within the household, and thus tested the ability of the model to describe the distribution of the number of cases within infected households. We define the subset $\Omega' \subset \Omega$ of pairs $\left\{m,n\right\}$ such that $m \geq 1$. The deviance is

$$\mathrm{Dev}_{\mathrm{IH}} = 2\sum_{\left\{m,n\right\}\in\Omega'}\left[k_{(m,n)}\left(\ln\left(\frac{O_m^n}{\sum_{s=1}^{n} O_s^n}\right) - \ln\left(\frac{P_m^n}{\sum_{s=1}^{n} P_s^n}\right)\right)\right] \qquad [22]$$

Both these deviance measures are included in Table S3. In comparing different models, it is apparent that prior immunity provides model fits which optimally describe the household attack rate, but that the distribution of cases within a household is best described by models with misreporting.

S7. Results of the model fitting and model comparison

Model	Δ AIC_c	AIC_c	K	Q	β	k	α	Q_{prior}	p_{asx}	p_{pr}	p_{com}	$SITP_3$	$SITP_9$	Dev	Dev_{IH}	Dev_{HAR}	†
PVS	0.00	173.16	5	0.80	0.37	0.94	0.35	0.88				0.20	0.15	162.45	154.08	29.30	X
PS	22.91	196.07	4	0.80	0.33		0.33	0.88				0.20	0.15	187.60	179.92	27.56	X
VS	24.35	197.51	4	0.80	0.20	0.76		0.88				0.17	0.17	189.04	180.49	26.51	X
S	57.95	231.11	3	0.80	0.18			0.89				0.16	0.16	224.83	214.94	25.97	X
PV	58.28	231.44	4	0.85	0.42	1.00	0.52					0.19	0.12	222.97	152.43	97.57	X
P	115.84	289.00	3	0.86	0.37		0.49					0.19	0.12	282.72	211.34	98.33	X
V	115.84	289.00	3	0.85	0.17	0.86						0.14	0.14	282.72	211.34	98.33	X
ReedFrost	188.35	361.51	2	0.85	0.15							0.14	0.14	357.37	286.90	97.60	X
PVAS	2.30	175.46	6	0.80	0.37	0.94	0.35	0.88	0.00			0.20	0.15	162.45	154.08	29.30	
PVXS	2.30	175.46	6	0.80	0.37	0.94	0.35	0.88		0.00		0.20	0.15	162.45	154.08	29.30	
PVAXS	4.65	177.81	7	0.80	0.37	0.94	0.35	0.88	0.00	0.00		0.20	0.15	162.45	154.08	29.30	
PAS	25.16	198.32	5	0.80	0.33		0.33	0.88	0.00			0.20	0.15	187.60	179.92	27.56	
PXS	25.16	198.32	5	0.80	0.33		0.33	0.88		0.00		0.20	0.15	187.60	179.92	27.56	
VXS	25.20	198.36	5	0.79	0.23	0.65		0.88		0.06		0.18	0.18	187.64	179.41	26.53	X
VAS	26.59	199.75	5	0.80	0.20	0.76		0.88	0.00			0.17	0.17	189.04	180.49	26.51	
PAXS	27.45	200.61	6	0.80	0.33		0.33	0.88	0.00	0.00		0.20	0.15	187.60	179.92	27.56	
VAXS	27.50	200.66	6	0.79	0.23	0.65		0.88	0.00	0.06		0.18	0.18	187.64	179.41	26.53	
AS	60.14	233.30	4	0.80	0.18			0.89	0.00			0.16	0.16	224.83	214.94	25.97	
XS	60.14	233.30	4	0.80	0.18			0.89		0.00		0.16	0.16	224.83	214.94	25.97	
PVA	60.52	233.68	5	0.85	0.42	1.00	0.52		0.00			0.19	0.12	222.97	152.43	97.57	
PVX	60.52	233.68	5	0.85	0.42	1.00	0.52			0.00		0.19	0.12	222.97	152.43	97.57	
AXS	62.39	235.55	5	0.80	0.18			0.89	0.00	0.00		0.16	0.16	224.83	214.94	25.97	
PVAX	62.82	235.98	6	0.85	0.42	1.00	0.52		0.00	0.00		0.19	0.12	222.97	152.43	97.57	
VX	111.04	284.20	4	0.83	0.22	0.69				0.12		0.18	0.18	275.73	204.61	98.08	X
PA	111.40	284.56	4	0.86	0.37		0.49		0.00			0.19	0.12	276.09	204.92	98.38	
PX	111.40	284.56	4	0.86	0.37		0.49			0.00		0.19	0.12	276.09	204.92	98.38	
VAX	113.28	286.44	5	0.83	0.22	0.68			0.00	0.12		0.18	0.18	275.73	204.57	98.13	
PAX	113.64	286.80	5	0.86	0.37		0.49		0.00	0.00		0.19	0.12	276.09	204.92	98.38	
VA	118.03	291.19	4	0.85	0.17	0.86			0.00			0.14	0.14	282.72	211.34	98.33	
A	190.49	363.65	3	0.85	0.15				0.00			0.14	0.14	357.37	286.90	97.60	
X	190.49	363.65	3	0.85	0.15					0.00		0.14	0.14	357.37	286.90	97.60	
AX	192.68	365.84	4	0.85	0.15				0.00	0.00		0.14	0.14	357.37	286.90	97.60	
PVRS	-13.30	159.86	6	0.75	0.38	0.29	0.45	0.94			0.78	0.16	0.11	146.85	134.66	34.09	X
PVR	-12.53	160.63	5	0.77	0.39	0.27	0.55				0.74	0.15	0.09	149.92	136.67	36.08	X
PVRAS	-10.94	162.22	7	0.75	0.38	0.29	0.45	0.94	0.00		0.78	0.16	0.11	146.86	134.67	34.11	
PVRXS	-10.94	162.22	7	0.75	0.38	0.29	0.45	0.94		0.00	0.78	0.16	0.11	146.86	134.67	34.11	
PVRA	-10.23	162.93	6	0.77	0.39	0.27	0.55		0.00		0.74	0.15	0.09	149.92	136.67	36.08	X
PVRX	-10.23	162.93	6	0.77	0.39	0.27	0.55			0.00	0.74	0.15	0.09	149.92	136.67	36.08	X
PVRAXS	-8.53	164.63	8	0.75	0.39	0.29	0.45	0.94	0.00	0.00	0.78	0.16	0.11	146.85	134.63	34.12	
PVRAX	-7.88	165.28	7	0.77	0.39	0.27	0.55		0.00	0.00	0.74	0.15	0.09	149.92	136.67	36.07	
VRS	1.80	174.96	5	0.74	0.18	0.25		0.92			0.76	0.13	0.13	164.24	150.97	33.47	
VRXS	3.38	176.54	6	0.73	0.20	0.23		0.92		0.03	0.77	0.14	0.14	163.52	150.66	33.04	
VRAS	4.10	177.26	6	0.74	0.18	0.25		0.92	0.00		0.76	0.13	0.13	164.24	150.97	33.47	
VRAXS	5.71	178.87	7	0.73	0.21	0.23		0.92	0.00	0.03	0.77	0.14	0.14	163.51	150.50	33.14	
VRX	7.89	181.05	5	0.72	0.24	0.16				0.12	0.72	0.13	0.13	170.33	157.77	34.81	X
VRAX	10.19	183.35	6	0.72	0.23	0.16			0.00	0.12	0.72	0.13	0.13	170.34	157.78	34.82	
VR	12.25	185.41	4	0.76	0.15	0.23					0.73	0.11	0.11	176.94	164.37	34.97	
VRA	14.12	187.28	5	0.75	0.17	0.20			0.03		0.72	0.11	0.11	176.56	164.33	34.55	X
PRS	25.16	198.32	5	0.80	0.33		0.33	0.88			1.00	0.20	0.15	187.60	179.92	27.56	
PRAS	27.45	200.61	6	0.80	0.33		0.33	0.88	0.00		1.00	0.20	0.15	187.60	179.92	27.56	
PRXS	27.45	200.61	6	0.80	0.33		0.33	0.88		0.00	1.00	0.20	0.15	187.60	179.92	27.56	
PRAXS	29.81	202.97	7	0.80	0.33		0.33	0.88	0.00	0.00	1.00	0.20	0.15	187.60	179.92	27.57	
RS	60.14	233.30	4	0.80	0.18			0.89			1.00	0.16	0.16	224.83	214.94	25.97	
RAS	62.39	235.55	5	0.80	0.18			0.89	0.00		1.00	0.16	0.16	224.83	214.94	25.97	
RXS	62.39	235.55	5	0.80	0.18			0.89		0.00	1.00	0.16	0.16	224.83	214.94	25.97	
RAXS	64.69	237.85	6	0.80	0.18			0.89	0.00	0.00	1.00	0.16	0.16	224.83	214.94	25.97	
PR	104.48	277.64	4	0.83	0.34		0.48				0.91	0.18	0.11	269.17	222.40	72.83	
PRA	106.72	279.88	5	0.83	0.34		0.48		0.00		0.91	0.18	0.11	269.17	222.40	72.82	
PRX	106.72	279.88	5	0.83	0.34		0.48			0.00	0.91	0.18	0.11	269.17	222.40	72.82	
PRAX	109.02	282.18	6	0.83	0.34		0.48		0.00	0.00	0.91	0.18	0.11	269.17	222.40	72.83	
R	169.16	342.32	3	0.82	0.14						0.86	0.13	0.13	336.04	303.68	57.51	
RA	171.35	344.51	4	0.82	0.14				0.00		0.86	0.13	0.13	336.04	303.68	57.51	
RX	171.35	344.51	4	0.82	0.14					0.00	0.86	0.13	0.13	336.04	303.68	57.51	
RAX	173.60	346.76	5	0.82	0.14				0.00	0.00	0.86	0.13	0.13	336.04	303.68	57.51	

Table S3 – The table summarizes the goodness of fit for all of the possible extensions of the basic Reed-Frost model. The table is divided into three sections: the first are the basic models considered in the main text; the second set are extensions of this first set which allow for the different types of asymptomatic infection; the third set allows for the possibility of systematic misreporting by household. The statistic

used to compare the models, $\Delta\mathrm{AIC}_c$*, is recorded relative to the first two sets of models, hence the negative values in the third set, representing models that fit better than baseline if one allows for systematic misreporting. Only fitted values of parameters are shown; default values for non-fitted parameters are* $k \to \infty$*,* $\alpha = 0$*,* $Q_{\mathrm{prior}} = 1$*,* $p_{\mathrm{asx}} = 0$*,* $p_{\mathrm{pr}} = 0$ *and* $p_{\mathrm{com}} = 1$*. †By default, the likelihood was optimized using the local optimizer 'FindMaximum' in Mathematica (18) with plausible initial guesses for the parameters. The fit of the marked models were checked using the global optimizer 'NMaximize' with default options. In no cases did this result in an improved fit, suggesting that 'FindMaximum' used an acceptable optimization algorithm for this task.*

S8. Confidence intervals

Confidence intervals were obtained from the univariate likelihood profiles. For the model analyzed in the main text, confidence intervals for were obtained for the **PVS** model (Table 1) ignoring the possible effect of asymptomatic infection. Confidence intervals for the parameter p_{asx} determining the proportion of infections that are asymptomatic and infectious were obtained for the **PVAS** model. Similarly, confidence intervals for p_{pr} were obtained for the **PVXS** model.

S9. Best fit parameters with 95% confidence intervals

9.1. The basic Reed-Frost model

Symbol	**Description**	**Best estimate and 95% Confidence Interval**
Q	Probability of not being exposed outside of the household at any stage in the fall epidemic wave	0.85 (0.85-0.86)
p	Susceptible-infectious transmission probability	0.14 (0.13-0.15)

Table S4 – best fit parameters and 95% confidence intervals for the basic Reed-Frost model

9.2. The PVS model

Symbol	Description	Best estimate and 95% Confidence Interval
Q	Probability of not being exposed outside of the household at any stage in the fall epidemic wave	0.80 (0.78-0.82)
β	Mean transmission parameter for within-household transmission	0.37 (0.29-0.48)
α	Coefficient for declining transmission as function of household size	0.35 (0.22-0.49)
k	Shape coefficient for variable individual infectiousness	0.93 (0.59-1.72)
Q_{prior}	Probability of not being exposed outside the household during the spring wave of the epidemic	0.88 (0.85-0.91)
p_{pr}	Proportion of individuals who are asymptomatic and not infectious†	0.00 (0.00-0.06)
p_{asx}	Proportion of individuals who are asymptomatic and fully infectious†	0.00 (0.00-0.03)

†parameters only included in an extended model.

Table S5 – best fit parameters and 95% confidence intervals for the ***PVS*** *model*

S10. Comparison of the model fit for several different models

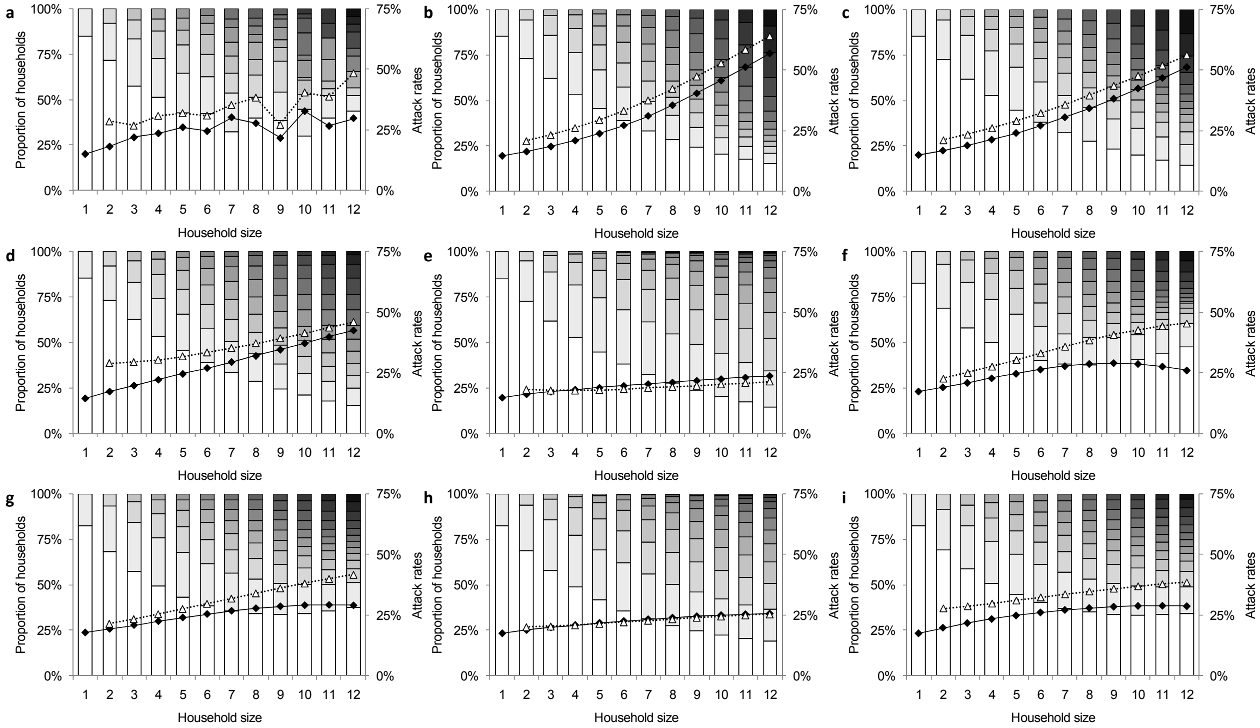

Figure S 3 – Plots showing the best fitting model for all the model variants included in the first part of Table S3 using the same display scheme as in Figure 1 and Figure S1. The panels ordered by increasing goodness of fit: ***a****, data from Baltimore;* ***b****, Reed Frost model;* ***c****, model* ***V****;* ***d****, model* ***P****;* ***e****, model* ***PV****;* ***f****, model* ***S****;* ***g****, model* ***VS****;* ***h****, model* ***PS*** *and* ***i****, model* ***PVS****.*

S11. Alternative visual representations of the results

The main results from the best fit models are presented in Figures 1 and 2. Alternative graphs give some additional insights into the nature of the fit.

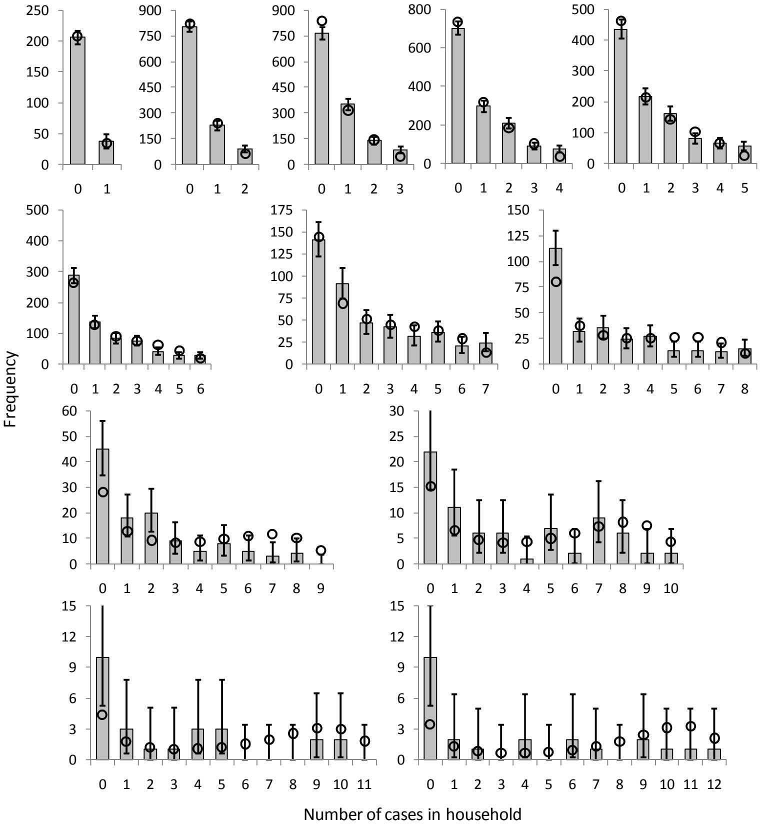

Figure S4 – the distribution of cases within households of different sizes in Baltimore (bars) and best fit Reed-Frost model (circles). Exact Binomial 95% confidence intervals are plotted for the cases in household as an indication of stochastic natural variation in random re-samples, included for indication only since goodness of fit is assessed by a full multinomial likelihood.

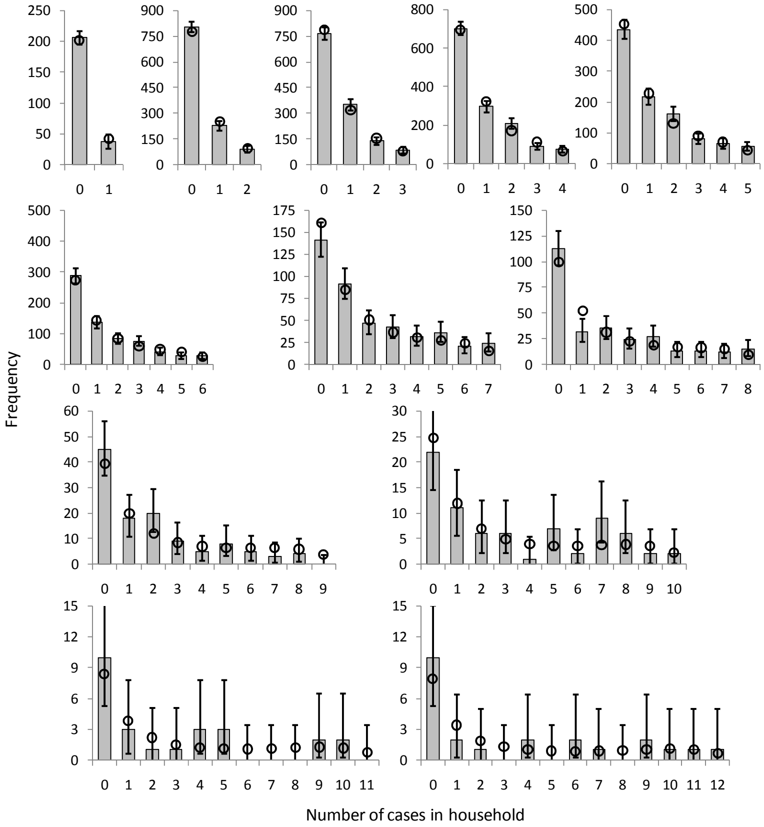

*Figure S5 – as Figure S4 –, but with the circles representing the prediction of the better fitting **PVS** model instead. The **PVR** model is not included as the difference in goodness of fit relative to the **PVS** model is not discernible to us using these graphs.*

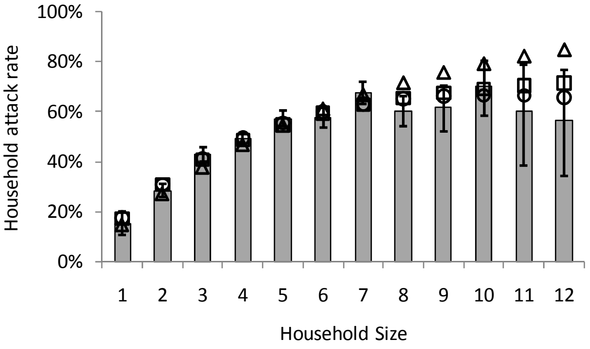

*Figure S6 –The household attack rate: the proportion of households in Baltimore experiencing at least one case (bars with exact Binomial 95% confidence intervals) and predictions of the best fit Reed-Frost model (triangles), **PVS** model (circles) and **PVR** model (squares).*

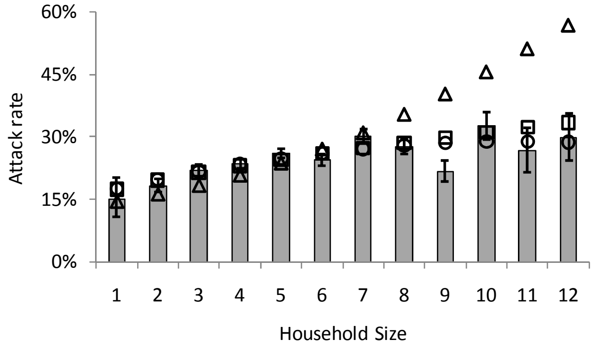

Figure S7 – The individual attack rate: the proportion of individuals with reported infection in Baltimore, stratified by size of household. Symbols are as in Figure S6 .

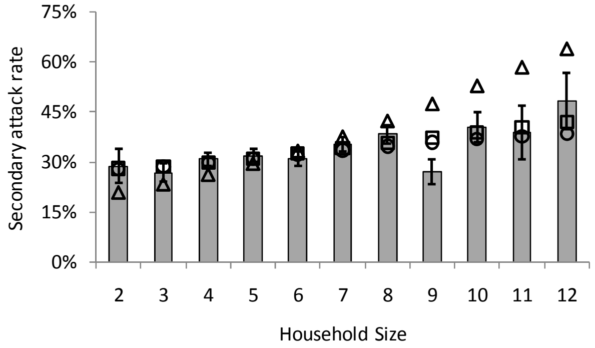

Figure S8 – The secondary attack rate: the proportion of secondary individuals with reported infection in households that experienced at least one case in Baltimore, stratified by size of household. Symbols are as in Figure S6 .

S12. Comparison of the epidemics in Baltimore and in Frederick

The **PVS** model was fit to the data from Frederick to obtain a comparison between different epidemics.

Symbol	Description	Baltimore	Frederick
Q	Probability of not being exposed outside of the household at any stage in the fall epidemic wave	0.80 (0.78-0.82)	0.80 (0.75-0.85)
β	Mean transmission parameter for within-household transmission	0.37 (0.29-0.48)	0.51 (0.25-1.00)
α	Coefficient for declining transmission as function of household size	0.35 (0.22-0.49)	0.43 (0.05-0.80)
k	Shape coefficient for variable individual infectiousness	0.94 (0.59-1.72)	0.94, fixed*
Q_{prior}	Probability of not being exposed outside the household during the spring wave of the epidemic	0.88 (0.85-0.91)	0.95 (0.87-1.00)

*Table S6 – Comparison of estimated parameters for the canvasses carried out in Baltimore and in Frederick (Document S1). *The shape parameter k was kept fixed to the best fit value for the Baltimore data as the sample size in Frederick was too small to estimate this reliably. If k was fitted as a parameter, the confidence interval was* $\left[1.68,\infty\right]$*, with the best fit obtained when* $k \to \infty$ *corresponding to constant infectiousness.*

The estimated parameters were very similar. Due to the smaller sample size, it was not possible to obtain information on the shape of the offspring distribution in Frederick, and so this was assumed to be similar to that estimated in Baltimore. The main difference between the epidemics was the significantly lower estimate of the magnitude of the inferred prior spring epidemic. This thus leads us to hypothesise that the higher overall attack rate in Frederick (32.1% versus 23.3% in Baltimore) may have been due to lower levels of prior immunity in that population.

S13. Comparison of the models with and without misreporting

As can be seen above, models which include systematic misreporting by household (i.e. models including **R** with $p_{\text{com}} \leq 1$) can fit the data better than models which do not include this. The overall best fit model is model **PVRS**, but this model suffers from parameter identifiability issues, due to a trade-off between the effect of **S** and **R** in the final distribution of cases in households. In the main text, we place most emphasis on model **PVS** which assumes no misreporting, while emphasizing which conclusions are or aren't robust to inclusion of some forms of systematic misreporting. This choice is motivated by the fact that models of misreporting can be somewhat arbitrary, and that our choice summarised by equation [11] might be one of many. The strengths of the conclusions of this analysis are dependent on the reliability of the original survey. We focus on model **PVR** which is a form of sensitivity analysis to assumptions about misreporting.

The principal effects of including the effects of possible misreporting were to reduce the estimated magnitude of prior immunity ($Q_{\text{prior}} = 0.88 \to 0.94$ in **PVRS**) and to increase the estimated degree of individual variability in infectiousness ($k = 0.94 \to 0.29$ in **PVR**). From the goodness of fit statistics presented in Table S3, we see that the **PVR** model tends to fit better than the **PVS** model to data within infected households, but less well to data on the proportions of households which actually get infected. A full comparison is included in ***Table S7.***

Symbol	**Description**	**Best estimate and 95% Confidence Interval**	**Estimates for PVS model for comparison**
Q	Probability of not being exposed outside of the household at any stage in the fall epidemic wave	0.77 (0.75-0.79)	0.80 (0.78-0.82)
β	Mean transmission parameter for within-household transmission	0.39 (0.28-0.57)	0.37 (0.29-0.48)
α	Coefficient for declining transmission as function of household size	0.54 (0.35-0.75)	0.35 (0.22-0.49)
k	Shape coefficient for variable individual infectiousness	0.27 (0.19-0.39)	0.94 (0.59-1.72)
p_{nc}	Probability of a household complying with the survey (otherwise record zero cases)	0.74 (0.71-0.79)	NA
p_{pr}	Proportion of individuals who are asymptomatic and not infectious†	0.00 (0.00-0.10)	0.00 (0.00-0.06)
p_{asx}	Proportion of individuals who are asymptomatic and fully infectious†	0.00 (0.00-0.07)	0.00 (0.00-0.03)

Table S7 – Best fit parameters for the **PVR** *model with systematic misreporting, and comparison with the best fit* ***PVS*** *model.*

S14. Exploring the dependence of infectiousness on household size

The statistical inference provided strong support for a dependence of infectiousness on household size. The model assumed a functional relationship between the transmission hazard and household size of the form $B_n = \beta / n^{\alpha}$ where β is drawn from a distribution. Given this distribution for the parameter β, the mean susceptible-infectious transmission probability is given by $\phi_n(1)$, where the moment generating function ϕ is defined in equation [6]. This is plotted in Fig. 2b for the best fit model.

The choice of the assumed functional dependence for infectiousness on household size, namely $B_n = \beta / n^{\alpha}$, was based on earlier work (13, 14). To explore the validity of this functional relationship, we considered an alternative non-parametric model where each B_n was considered as an independent parameter, also plotted in Fig 2b. To avoid the problem of parameter over-specification when considering confidence intervals for the parameters B_n , these were derived only within the subspace of the whole parameter space where all the other parameters were kept fixed at their best fit values for the **PVS** model (i.e. those given in Table 1). The agreement between the functional curve and the non-parametric model seen in Fig. 2b suggests that the choice of function was acceptable. It would be interesting in future work to explore the alternative choice given by $B_n = \beta / (n-1)^\alpha$ which provides an alternative interpolation between frequency and density dependent transmission (when $\alpha = 0$ and $\alpha = 1$ respectively).

S15. Estimation of the susceptible-infectious transmission probability

The susceptible-infectious transmission probability is defined as the cumulative probability that an infectious individual infects a susceptible individual in a household (assuming that the susceptible individual remains susceptible for the duration of the infectiousness of his or her infectious contact). In general, it depends on the household size, and is given by $\mathrm{SITP}_n = 1 - \phi_n(1)$ where $\phi_n(x)$ is the moment generating function of infectiousness defined earlier. In general this will depend on household size, and we thus define the susceptible-infectious transmission probability for a randomly chosen infected individual as

$$\mathrm{SITP} = 1 - \sum_{n=0}^{n_{\max}} \phi_n(1) f[n] \qquad [23]$$

where $f[n]$ is the size-biased household size distribution defined earlier. We derived the best estimate for this by substituting maximum likelihood parameters into equation [23]. For the Reed-Frost model, $\phi_n(1) = \exp(-\beta)$ and so confidence intervals were obtained from the confidence intervals for β . For other models, we obtained conservative confidence intervals by substituting lower and upper bounds of confidence intervals of the non-parametric estimates of B_n described in the previous section into equation [23].

S16. Exploring the impact of large households on inferred parameters

Based on the summary statistics used to describe the goodness of fit, it is reasonable to ask whether the model inferences are driven by specific differences in the outcome of the epidemic in large households. To test this, we analyzed the epidemic within the subset of the sample restricted to households of size equal to or less than six. The results were not appreciably different (***Table S8***), indicating that the inferred results were not specifically driven by the outcomes observed in large households.

Model	Δ AIC$_c$	AIC$_c$	K	Q	β	k	α	Q$_{prior}$	p$_{asx}$	p$_{pr}$	p$_{com}$	SITP$_3$	SITP$_9$	Dev	Dev $_{IH}$	Dev $_{HAR}$	†
PVS	0.00	173.16	5	0.80	0.37	0.94	0.35	0.88				0.20	0.15	162.45	154.08	29.30	X
PS	22.91	196.07	4	0.80	0.33		0.33	0.88				0.20	0.15	187.60	179.92	27.56	X
VS	24.35	197.51	4	0.80	0.20	0.76		0.88				0.17	0.17	189.04	180.49	26.51	X
S	57.95	231.11	3	0.80	0.18			0.89				0.16	0.16	224.83	214.94	25.97	X
PV	58.28	231.44	4	0.85	0.42	1.00	0.52					0.19	0.12	222.97	152.43	97.57	X
P	115.84	289.00	3	0.86	0.37		0.49					0.19	0.12	282.72	211.34	98.33	X
V	115.84	289.00	3	0.85	0.17	0.86						0.14	0.14	282.72	211.34	98.33	X
ReedFrost	188.35	361.51	2	0.85	0.15							0.14	0.14	357.37	286.90	97.60	X

Model	Δ AIC$_c$	AIC$_c$	K	Q	β	k	α	Q$_{prior}$	p$_{asx}$	p$_{pr}$	p$_{com}$	SITP$_3$	SITP$_9$	Dev	Dev $_{IH}$	Dev $_{HAR}$	†
PVS	0.00	27.15	5	0.84	0.39	0.81	0.40	0.90				0.20	0.14	29.15	20.94	8.24	
PV	7.29	34.45	4	0.84	0.40	0.90	0.51					0.18	0.12	36.45	20.58	15.98	
VS	9.55	36.71	4	0.80	0.22	0.82		0.88				0.18	0.18	38.71	30.77	7.84	
PS	20.15	47.31	4	0.80	0.33		0.32	0.86				0.21	0.15	49.31	39.82	9.67	
V	24.61	51.77	3	0.84	0.19	0.88						0.16	0.16	53.77	37.77	16.07	
S	28.97	56.12	3	0.79	0.20			0.86				0.18	0.18	58.12	47.22	10.92	
P	38.93	66.08	3	0.85	0.34		0.46					0.19	0.12	68.08	52.02	16.41	
ReedFrost	58.77	85.93	2	0.85	0.17							0.16	0.16	87.93	72.03	16.20	

Table S8 – Best fit parameters for a range of models fitted to the data from Baltimore for all households (above) and for households of size less or equal to six (below). Table headings are as in Table S3.

S17. The offspring distribution in the large population limit

To compare our model of heterogeneous infection hazards to an earlier study of heterogeneous infectiousness in epidemic models (20), we consider the large population limit of our model of within-household transmission. Heterogeneity is assessed by the variation in the number of offspring (i.e. new infections) which each infected individual generates over the total course of infectiousness, called the offspring distribution (21). Extending our model to the large population limit, mixing is assumed homogeneous and random, and infectious individuals infect others with hazard h_E. This hazard is a drawn from a gamma distribution with mean R and shape parameter k, where R is the individual reproduction number and k is as defined in [5]. Given a hazard h_E, an infectious individual infects a number x of new individuals drawn from the Poisson distribution:

$$\mathrm{pr}\left(x \mid h_E\right) = \frac{\left(h_E\right)^x \exp\left(-h_E\right)}{\Gamma\left(1+x\right)} \qquad [24]$$

Averaging this distribution over the underlying distribution of hazards, the distribution of the number infected is given by a negative binomial distribution with mean R and shape parameter k, i.e.

$$\mathrm{pr}\left(x\right) = \frac{\Gamma\left(x+k\right)}{\Gamma\left(x+1\right)\Gamma\left(k\right)} \frac{k^x R^k}{\left(k+R\right)^{x+k}} \qquad [25]$$

In other words, the shape parameter k which measures the heterogeneity in infectiousness in the model of within-household transmission can be directly equated with the shape parameter k associated with a negative binomial offspring distribution which has been estimated in other epidemics and outbreaks (20).

S18. Two-group high-spreader/low-spreader model variants

To test the sensitivity of our analysis to our assumed parametric shape for the offspring distribution, we considered an alternative model consisting of a superposition of two Poisson distributions, corresponding to a sub-group of the population being more infectious than the rest. This assumption is quite general, but can be viewed in a special case as a two class children/adult model in the light of our discussion of age structure below.

In this model variant, there are two possible infection hazards $\beta_{\text{high}} / n^{\alpha}$ and $\beta_{\text{low}} / n^{\alpha}$ which arise with probability p_{high} and $1-p_{\text{high}}$ respectively. The moment generating function is then

$$\phi_{\text{2-group}}\left(x\right) = p_{\text{high}}\exp\left(-x\beta_{\text{high}} / n^{\alpha}\right) + \left(1-p_{\text{high}}\right)\exp\left(-x\beta_{\text{low}} / n^{\alpha}\right) \qquad [26]$$

The standard Reed-Frost model is recovered when $\alpha = 0$, $p_{\text{high}} = 0$ and $\beta_{\text{low}} = \beta$. We denote by **T** models which include this two-group structure, and consider models **PTS** and **PTR** as alternatives to **PVS** and **PVR**.

Symbol	Description	PTS	PVS	PTR	PVR
Q	Probability of not being exposed outside of the household at any stage in the fall epidemic wave	0.81	0.80	0.77	0.77
α	Coefficient for declining transmission as function of household size	0.38	0.35	0.53	0.54
Q_{prior}	Probability of not being exposed outside of the household in the spring wave epidemic	0.89	0.88	-	-
p_{nc}	Probability of a household complying with the survey (otherwise record zero cases)	-	-	0.76	0.74
β	Mean transmission parameter for within-household transmission	-	0.37	-	0.39
k	Shape coefficient for variable individual infectiousness	-	0.94	-	0.27
β_{high}	Transmission parameter for high infectiousness group	3.92	-	1.88	-
β_{low}	Transmission parameter for low infectiousness group	0.28	-	0.12	-
p_{high}	Proportion of individuals who are highly infectious	0.048	-	0.15	-
Dev	Deviance, a goodness of fit measure	141.00	162.45	152.74	149.92
ΔAIC_c	Comparative goodness of fit measure relative to **PVS** model	-19.15	0.00	-7.41	-12.53
SITP_3	The susceptible-infectious transmission probability in a household of size 3	0.20	0.20	0.15	0.15
SITP_9	The susceptible-infectious transmission probability in a household of size 9	0.15	0.15	0.10	0.09

Table S9 – Best fit parameters for the two-class PTS *and* PTR *models and* PVS *and* PVR *models presented for comparison.*

These supplementary sensitivity analyses highlight two aspects of our conclusions. First, our main conclusions regarding the low infectiousness and heterogeneity in offspring distribution are robust to consideration of different forms of the model. Second, that the relative support for either a model including a spring wave epidemic or a specific form of misreporting is dependent on the details of the model chosen. The **PTS** model is the best fitting of all tested, but relies on a somewhat artificial choice of model for the offspring distribution that may *a priori* be considered implausible.

S19. Variation in susceptibility

Most of our model variants are based on considering variation in individuals' infectiousness, but individuals also vary in their susceptibility to infection given exposure. While we do not include this explicitly in our models due to the computational challenge involved, we argue on qualitative grounds that models involving substantial variability in susceptibility are unlikely to be strongly supported by the Frost-Sydenstricker data.

At one extreme, model variants which include the asymptomatic uninfectious state (models including **P** state) can be also considered as models including a subset of people with zero susceptibility to infection (a protected class), and thus an extreme case of variable susceptibility. Almost all these models variants receive no support from the data (see estimates of p_{pr} in ***Table S3***).

More generally, the effect of variable susceptibility to infection is to decrease the variance of the distribution of the number of final cases in the household(8), and model variants which are favoured in the model comparison exercise are ones which increase the variance relative to the basic Reed-Frost model. Thus, we hypothesise that if we were to include variable susceptibility in our analysis, we would as a result obtain estimates with higher variability in infectiousness to compensate.

Some variability in susceptibility is likely for all infections, and for influenza, much interest focuses on age-dependence. We analyse some complementary data below which shows quite mild variation in susceptibility by age compared to studies of influenza in other epidemics.

S20. Age structure

One aspect of influenza transmission which has been considered important in many studies is the dependence of infection rates on age (see (14, 22-25) and references therein). Either because of mixing patterns, biology, or both, children are more susceptible to infection and may also be more infectious when infected. While we do not have sufficient data for a full age stratified model of the Frost-Sydenstricker study, some summary data exist (Document S1). The overall composition of the households by age group was recorded, albeit in an incomplete manner.

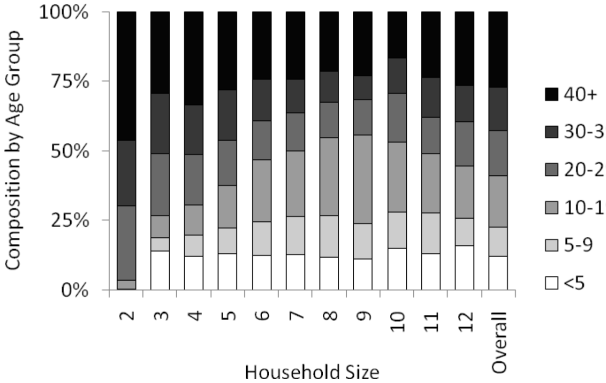

Figure S9 – the distribution of age classes within infected households, stratified by size of the household, and excluding the index case in the household.

We did not discern striking trends in the age distribution apart from the relative rarity of children in households of size 2, as might be expected. These were the households which experienced the highest susceptible-infectious transmission probabilities.

The documents also record the age structure in cases subsequent to the original index case in the household, from which some information on the age-dependent susceptibility can be gleaned.

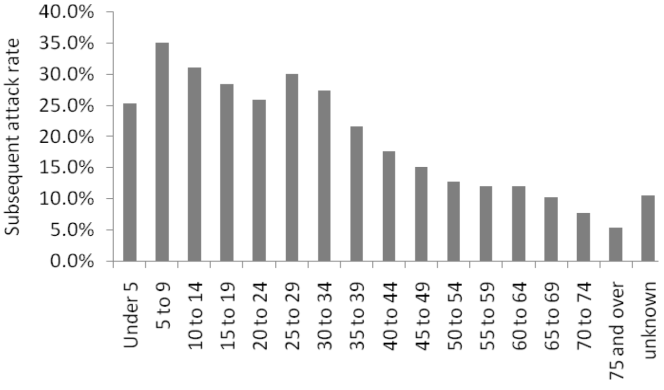

Figure S10 – the probability of ultimately being infected within a household when not an index case, stratified by age.

Age-dependent susceptibility does not appear to decline systematically until age 35. This is consistent with a hypothesis of some cross-protection arising from infection by viruses circulating between the 1847 and 1889 influenza pandemics (26).

To explore whether the inclusion of age structure could modify our results, we modify the two class model defined in section S18 to model a two group structure where we divide the population into more infectious young individuals (age 19 or less) making up 37.4% of the population (based on the incomplete data summarised in figure *Figure S9*), and the rest. We use **G** to denote model variants which include this type of age structure.

Symbol	**Description**	**PGS**	**PGVS**	**PVS**	**PGR**	**PGVR**	**PVR**
Q	Probability of not being exposed outside of the household at any stage in the fall epidemic wave	0.80	0.80	0.80	0.79	0.77	0.77
α	Coefficient for declining transmission as function of household size	0.36	0.33	0.35	0.61	0.54	0.54
Q_{prior}	Probability of not being exposed outside of the household in the spring wave epidemic	0.88	0.88	0.88	-	-	-
p_{nc}	Probability of a household complying with the survey (otherwise record zero cases)	-	-	-	0.80	0.75	0.74
β	Mean transmission parameter for within-household transmission	-	-	0.37	-	-	0.39
k	Shape coefficient for variable individual infectiousness	-	2.46	0.94	-	0.33	0.27
β_{high}	Transmission parameter for high infectiousness group	0.76	0.65	-	1.03	0.67	-
β_{low}	Transmission parameter for low infectiousness group	0.13	0.18	-	0.06	0.24	-
p_{high}	Proportion of individuals who are highly infectious	0.37*	0.37*	-	0.37*	0.37*	-
Dev	Deviance, a goodness of fit measure	167.13	160.52	162.45	176.74	149.73	149.92
ΔAIC_c	Comparative goodness of fit	6.99	2.73	0.00	16.59	-8.06	-12.53

	measure relative to **PVS** model					
SITP_3	The susceptible-infectious transmission probability in a household of size 3	0.20	0.20	0.20	0.17	0.15	0.15
SITP_9	The susceptible-infectious transmission probability in a household of size 9	0.15	0.15	0.15	0.10	0.09	0.09

Table S10 – Best fit parameters for the two-class age-stratified PGS/PGVS *and* PGR/PGVR *models and* PVS *and* PVR *models presented for comparison. *The proportion* p_{high} *is fixed to* 37.4%*, the proportion of under 20s estimated in the population.*

As for the two-class model considered in section S18, the main outcome of including age-structure in the model, seen in ***Table S10*** is to modify the offspring distribution. This particular modification is not statistically supported, which does not of course reject age-related variation in infectiousness, but rather indicates that the data we have do not enable us to detect it. Other parameters are not substantially modified by these changes in the offspring distribution.

S21. Time-series analysis

21.1. The instantaneous individual reproduction number

The instantaneous reproduction number $R(t)$ was introduced in (27) to estimate changes in the reproduction number over time from time-series of incident cases in an epidemic. It is one of a number of related methods developed to do this (28-30), all of which are equivalent in the case of an exponentially growing epidemic (31). The method used here is based on the simplest renewal equation for an epidemic (27), determined from the time series of incident cases I_t and with knowledge of the generation time distribution, denoted ω_t. The estimate is given by

$$R(t) = \frac{I_t}{\sum_{s=1}^{t} I_{t-s}\omega_s} \qquad [27]$$

The generation time distribution ω_t is defined as the distribution of times taken between a person being infected and infecting another (defined forwards in time and over all actual transmission events). More precisely, ω_t is the probability of such a time lying in the continuous interval $]t-1,t]$.

21.2. The instantaneous household reproduction number

The instantaneous household reproduction number $R^*(t)$ was also introduced in (27) to generalize the concept of reproduction in a way which explicitly accounts for household structure. It is defined as the number of households one newly infected household at time t may be expected to infect over the whole duration of the within-household outbreak, should conditions remain the same. It is defined by generalizing equation [27] to

$$R^*(t) = \frac{I_t}{\sum_{s=1}^{t} I_{t-s}\omega_s^*} \qquad [28]$$

The household generation time distribution ω_t^* is defined in (27) as the mean time taken for one household to infect another. In this case, it is estimated numerically based on simulating 100,000 household outbreaks using an algorithm analogous to that described in the supplementary information of (27), but adapted for the best-fitting model of within household outbreaks estimated here. Outbreaks are assumed to be started by an individual drawn randomly from the susceptible population; prior immunity from the spring wave is generated from the best-fitting distribution $F_l^{n,n}\left[Q_{\text{prior}}\right]$.

21.3. Dependence on generation time distribution

The generation time distribution is difficult to estimate in practice, and few estimates exist for influenza. We use the estimates of ω_t from adapted from reference (31) (as in (27)) based on seasonal influenza which is relatively robust and represents a relative consensus between variable published values (21). Estimates were also adapted from (14, 22) for sensitivity analysis.

For Choice 1 ((31), (27)), the mean generation time is 2.85 days with standard deviation 0.93 days. For Choice 2 (14), the mean generation time is 2.67 days with standard deviation 1.81 days. For Choice 3 (22), the mean generation time is 5.30 days with standard deviation 4.27 days.

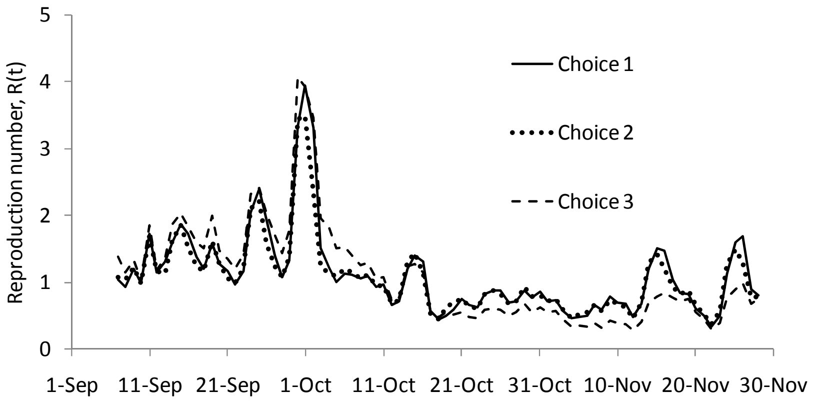

Figure S11 – comparison of estimates of the individual reproduction number $R(t)$ *for three different choices of the generation time distribution. The mean values of* $R(t)$*for the period up to 10-Oct are* 1.52 *for Choice 1 (used in the main text)(31),* 1.45*for Choice 2(14) and* 1.77 *for Choice 3(22). For the subsequent time-period (end of the epidemic wave), the values were* 0.80, 0.79 *and* 0.61 *respectively.*

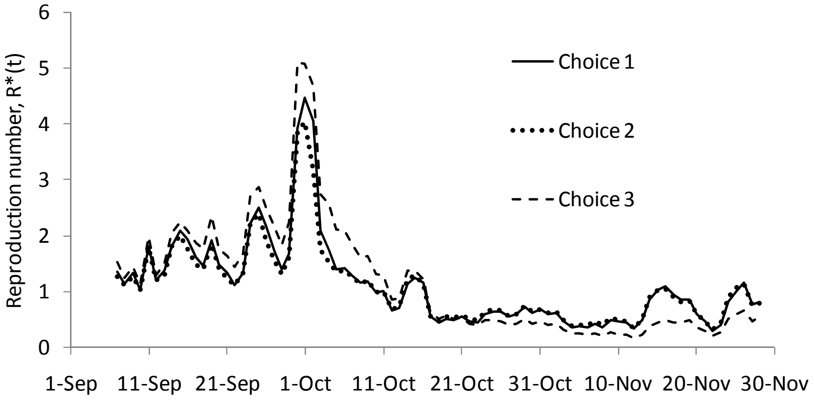

Figure S12 – comparison of estimates of the household reproduction number $R(t)$ *for three different choices of the generation time distribution. The mean values of* $R(t)$*for the period up to 10-Oct are* 1.79 *for Choice 1 (used in the main text)(31),* 1.67*for Choice 2(24) and* 2.17 *for Choice 3(22). For the subsequent time-period (end of the epidemic wave), the values were* 0.66, 0.67 *and* 0.49 *respectively.*

21.4. An improved estimator for the reproduction numbers

To move beyond point estimates of the reproduction number, we develop a likelihood-based method for estimating reproduction numbers. The renewal equation needs to be extended to include stochastic variation, so that the likelihood can be defined as the probability of observing the data given the model. A stochastic model equivalent to equation [27] is derived by assuming that daily incidence counts are distributed around their expectation according to a Poisson distribution, i.e.

$$I_t \sim \text{Poisson}\left(R_t \sum_{s=1}^{t} I_{t-s}\omega_s\right) \qquad [29]$$

A likelihood for the reproduction numbers is then given by

$$l(\{R_t\}) = const + \sum_t \left[I_t \ln\left(R_t \sum_{s=1}^{t} I_{t-s}\omega_s\right) - R_t \sum_{s=1}^{t} I_{t-s}\omega_s \right] \qquad [30]$$

Non-smoothed renewal-type estimates of the reproduction number from epidemic time-series typically suffer from heavy negative autocorrelation (as seen by large fluctuations in Figure S11 and S12). Improved stability of estimates is obtained by reducing the number of time points where R_t can change.

For the estimates presented in Figure 3, R_t was considered piecewise constant over 10 day intervals, while for the estimates presented in Table 1, R_t was assumed constant before and after the 10th October. In both cases estimates were derived by maximum likelihood using equation [30].

S22. Analysis of transmission of seasonal influenza in modern households

Two published studies have characterized the transmission of seasonal H3N2 influenza in households (13, 32, 33). To enable a model-consistent comparison with our results from the 1918 pandemic, we re-analyse the data from the Tecumseh (32, 33) and Epigrippe (13) studies here using our model.

These studies were different from the 1918 study, and different from each other, and thus required the model to be adapted and comparisons to be interpreted with caution.

		Number of persons in household, *n*			
Cases in household, *m*		**1**	**2**	**3**	**4**
	0	45	52	17	16
	1	18	11	4	4
	2		8	3	6
	3			5	0
	4				2

Table S11 - the distribution households according to size and number of influenza cases reported in one influenza season, based on the canvass in Tecumseh. These data are extracted from Table 4 in (33) and consist of all the households where all individuals have low influenza specific antibody titres prior to the study period. Thus in this case the probability of lack of prior immunity didn't need to be estimated, rather it could be assumed =0%. Similarly, this study measured seroconversion, not symptomatic infection and thus we did not need to account separately for asymptomatic cases.

		Number of persons in household, *n*			
Cases in household, *m*		**2**	**3**	**4**	**5**
	1	53	28	31	11
	2	49	24	22	11
	3		23	33	10
	4			18	7
	5				4

Table S12 - the distribution households according to size and number of influenza cases, based on the Epigrippe study. In this study households were followed to detect secondary infections for two weeks following the presentation of an index symptomatic influenza case in a general practice surgery (13).This study was based on symptoms in secondary cases – we assumed a single index case and that all secondary cases has manifested themselves over the two week follow-up period.

Because these studies were small, it was not possible to identify as many parameters as in the 1918 study. In analyzing both studies, we assume no prior immunity ($Q_{\text{prior}} = 1$). The shape parameter is assumed fixed to the best fit value in model PVS, i.e. $k = 0.94$.

To analyse the Tecumseh study data, the full final distribution without prior immunity, $F_m^{n,n}\left[Q\right]$ defined by equation [4], was used. The fit parameters were $\beta = 1.99$, $\alpha = 1.43$ and $Q = 0.83$.

The Epigrippe study follows an outbreak initiated by an index case, and we thus use the appropriate model distribution for outbreaks, i.e. G_m^n defined by

$$\binom{n}{k} = \sum_{m=0}^{k} \binom{n-1-m}{k-m} G_m^n \Big/ \phi\left(n-1-k\right)^{1+m} \quad \text{for} \quad k = 0,\ldots,n-1 \qquad [31]$$

(from references (10, 11)). In this case the best fit parameters were $\beta = 2.43$ and $\alpha = 1.30$. Note that there may have been some selection bias for more severe symptoms and thus higher infectiousness in index cases (14).

For a more informative comparison with the results from the Frost 1918 study, we compare susceptible-infectious transmission probabilities stratified by household size, using the non-parametric method described in section S14. The results are summarised in Figure 2A.This shows that the estimates of susceptible-infectious transmission probabilities for the Frost study in 1918 are lower than comparable estimates for more recent studies of seasonal influenza.

It also highlights that the dependence of the susceptible-infectious transmission probability on household size is less marked in 1918 than in these more recent studies. Whether this is due to differences in study design, differences in the virus, or secular trends in household mixing patterns is unclear.

S23. Analysis of H1N1 pandemic influenza transmission in households in 2009

A recent study has characterized transmission of pandemic H1N1 virus in households in the USA (34). Here, we re-analyse these data using our model. The final-size data are summarized below.

		Number of persons in household, *n*				
		2	**3**	**4**	**5**	**6**
Cases in household, *m*	**1**	28	34	52	31	11
	2	11	9	13	9	4
	3		4	2	4	0
	4			2	1	1
	5				0	0
	6					0

Table S13 - the distribution households according to size and number of influenza cases reported in a period +/- 7 days centred around the report of an index case. Cases were defined based on reports of acute respiratory infection.

The study was analyzed using the same outbreak model as used for the Epigrippe study of seasonal influenza (previous section), and we also found that the study was too small to reliably identify more than the basic parameters which determine infectiousness. We assumed that the household outbreak was complete within the 7 day follow-up, thus ignoring censoring effects in the data. This approximation seemed reasonable given the short generation time estimates and low secondary attack rate estimates for this infection (34).

The best fit parameters were $\beta = 0.92$ with 95% confidence interval $\left(0.31 - 2.99\right)$, $\alpha = 1.44\ \left(0.72 - 2.31\right)$ and $k = 2.56\ \left(0.42 - \infty\right)$. We also estimated susceptible-infectious transmission probabilities stratified by household size (Figure 2A).

Because of evidence of both prior cross-reactive immunity and asymptomatic infection having played a role in the 2009 pandemic (35), we examined the sensitivity of our estimates to different assumptions about asymptomatic infections and prior immunity in these households.

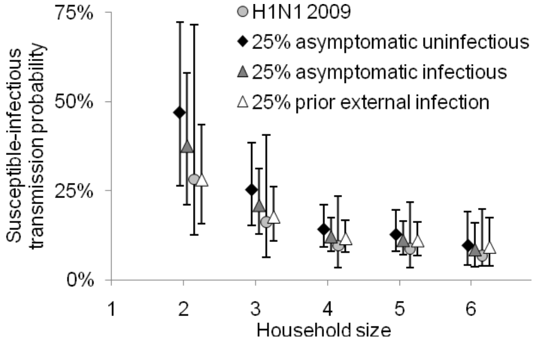

Figure S13 – sensitivity analysis of estimates to different assumptions about parameters that could not be estimated for this study. In the main scenario there was no prior immunity $\left(Q_{\text{prior}} = 1\right)$ *or asymptomatic infections* $p_{\text{asx}} = 0$ *and* $p_{\text{pr}} = 0$. *We then considered three alternative scenarios where 1)* $p_{\text{pr}} = 0.25$, $p_{\text{asx}} = 0$ *and* $Q_{\text{prior}} = 1$, *2)* $p_{\text{pr}} = 0$, $p_{\text{asx}} = 0.25$ *and* $Q_{\text{prior}} = 1$ *and 3)* $p_{\text{pr}} = 0$, $p_{\text{asx}} = 0$ *and* $Q_{\text{prior}} = 0.75$. *The overall susceptible-infectious transmission probability (averaged over the size-biased distribution of household sizes) was main:* $11.9\%\ \left(7.1\% - 18.2\%\right)$, *1):* $18.4\%\ \left(10.9\% - 28.3\%\right)$, *2)* $15.4\%\ \left(9.3\% - 23.3\%\right)$ *and 3)* $13.8\%\ \left(8.3\% - 20.8\%\right)$.

S24. References

1. Daniel TD. Wade Hampton Frost, Pioneer Epidemiologist 1880-1938: Up to the Mountain. Rochester: University of Rochester Press, 2004.
2. Frost WH, Sydenstricker E. Influenza in Maryland: preliminary statistics of certain localities. *Public Health Rep.* 1919:491-504.
3. Sydenstricker E. The incidence of influenza among persons of different economic status during the epidemic of 1918. *Public Health Rep.* 1931;121:191-204.
4. Oxford JS, Lambkin R, Sefton A, et al. A hypothesis: the conjunction of soldiers, gas, pigs, ducks, geese and horses in northern France during the Great War provided the conditions for the emergence of the "Spanish" influenza pandemic of 1918-1919. *Vaccine* 2005;23:940-5.
5. Department of Public Safety. Annual Report of the Sub-Department of Health to the Mayor and City Council of Baltimore for the Fiscal Year Ended Dec. 31, 1918. Baltimore: King Bros. City Printers, 1920.
6. Ministry of Health. Report on the Pandemic of Influenza 1918-19. London: His Majesty's Stationary Office, 1920.
7. Frost WH. Some conceptions of epidemics in general. *Am. J. Epidemiol.* 1976;103:141-151.
8. Ludwig D. Final size distributions for epidemics. *Math. Biosci.* 1975;23:33-46.
9. Pellis L, Ferguson NM, Fraser C. The relationship between real-time and discrete-generation models of epidemic spread. *Math. Biosci.* 2008;216:63-70.
10. Addy CL, Longini IM, Haber M. A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics* 1991;47:961-974.
11. Ball F, Mollison D, Scalia-Tomba G. Epidemics with two levels of mixing. *Ann. Appl. Probab.* 1997;7:46-89.
12. Andersson H, Britton T. Stochastic Epidemic Models and Their Statistical Analysis. New York: Springer, 2000.
13. Cauchemez S, Carrat F, Viboud C, et al. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Stat. Med.* 2004;23:3469-3487.
14. Ferguson NM, Cummings DAT, Cauchemez S, et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 2005;437:209-214.
15. Olson DR, Simonsen L, Edelson PJ, et al. Epidemiological evidence of an early wave of the 1918 influenza pandemic in New York City. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102:11059-63.
16. Andreasen V, Viboud C, Simonsen L. Epidemiologic characterization of the 1918 influenza pandemic summer wave in Copenhagen: implications for pandemic control strategies. *J. Infect. Dis.* 2008;197:270-8.
17. Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. *Nature* 2003;422:428-33.
18. Wolfram Research Inc. Mathematica, Version 7.0. Champaign, IL, 2009.
19. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer, 2002.
20. Lloyd-Smith JO, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. *Nature* 2005;438:355-359.
21. Grassly NC, Fraser C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* 2008;6:477-487.
22. Longini IM, Nizam A, Xu SF, et al. Containing pandemic influenza at the source. *Science* 2005;309:1083-1087.
23. Cauchemez S, Valleron AJ, Boelle PY, et al. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature* 2008;452:750-4.

24. Ferguson NM, Cummings DAT, Fraser C, et al. Strategies for mitigating an influenza pandemic. *Nature* 2006;442:448-452.
25. Germann TC, Kadau K, Longini IM, et al. Mitigation strategies for pandemic influenza in the United States. *Proc. Natl. Acad. Sci. U.S.A.* 2006;103:5935-5940.
26. Taubenberger JK, Morens DM, Fauci AS. The next influenza pandemic: can it be predicted? *JAMA* 2007;297:2025-7.
27. Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE* 2007;2:e758.
28. Cauchemez S, Boelle PY, Thomas G, et al. Estimating in real time the efficacy of measures to control emerging communicable diseases. *Am. J. Epidemiol.* 2006;164:591-597.
29. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* 2004;160:509-16.
30. White LF, Pagano M. Transmissibility of the influenza virus in the 1918 pandemic. *PLoS ONE* 2008;3:e1498.
31. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. Lond. B Biol. Sci.* 2007;274:599-604.
32. Longini IM, Koopman JS, Monto AS, et al. Estimating Household and Community Transmission Parameters for Influenza. *Am. J. Epidemiol.* 1982;115:736-751.
33. Longini IM, Koopman JS, Haber M, et al. Statistical Inference for Infectious Diseases: Risk Specific Household and Community Transmission Parameters. *Am. J. Epidemiol.* 1988;128:845-859.
34. 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;361:2619-2627.
35. Miller E, Hoschler K, Hardelid P, et al. Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. *Lancet* 2010;375:1100-1108.