Supplementary Appendix

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

Supplement to: Lessler J, Reich NG, Cummings DAT, et al. Outbreak of 2009 pandemic influenza A (H1N1) at a New York City school. N Engl J Med 2009;361:2628-36.

Supplementary material for
Outbreak of 2009 H1N1 at a New York City School

DOHMH Swine Influenza Investigation Team, Justin Lessler, Nicholas G. Reich and Derek A.T. Cummings

This supplement describes the statistical methods used in the analysis of the natural history of a novel swine-origin influenza A (2009 H1N1) virus and lists the additional New York City Department of Health and Mental Hygiene Staff who aided with the outbreak response.

1 Incubation period

The infection time and symptom onset time for cases are needed to estimate the distribution of the incubation period. Often one or both of these times are not observed exactly and are instead observed to lie within an interval of time. An observation with both observed to be within an interval is called a doubly interval-censored observation. All of the incubation period data collected in this outbreak was treated as doubly interval-censored.

For all cases, the symptom onset time was reported as a single calendar day. The exposure history for each case report was examined individually. Cases were classified into one of three categories: (1) a student or employee at High School A, (2) a household member of a student or employee at High School A, or (3) a community contact of a High School A student or employee. We developed a set of rules for determining exposure times that was specific to each of these three classes of cases.

For students or employees at High School A, Monday, April 20th was fixed as the first possible day of infection. This was the first day that presumably infectious students who were returning from Mexico were attending classes at the school. Unless specific information from a survey indicated otherwise, the earliest possible infection time for all cases in this group was chosen to be 0600h on Monday, April 20th and the latest possible infection time was chosen to be their time of symptom onset. For the one confirmed case who had traveled to Mexico, the earliest exposure date was set as the arrival date in Mexico.

For a household member of a student or employee at High School A, the exposure times were determined by the onset date of the index case in the household. Based on published patterns of viral load of influenza across the course of infection [1], we assumed that identified contacts may have been infectious on the day prior to their symptom onset and this was fixed as the earliest possible day of infection. The latest possible infection time was chosen as the time of symptom onset (i.e. we assume individuals could have been infected up until the time they became symptomatic). Again, any specific information provided in a survey or interview was considered on a case-by-case basis in determining these times.

For community-based contacts of a High School A student or employee, the possible infection times were often reported in surveys or interviews conducted with the cases in the days following the outbreak. When contact with a confirmed case from High School A was reported to have occurred on a specific day, that day was chosen as the only window of possible infection. When the contact time with a confirmed case from High

School A was not given precisely, the window of possible exposures was calculated in the same manner as for a student or employee at High School A.

These rules generated observations of the form (E_L, E_R, S_L, S_R) where E_L and E_R are the left and right endpoints on the possible exposure interval and S_L and S_R are the same for the symptom onset interval. The likelihood for a single observation of this form is

$$L(\theta; E_L, E_R, S_L, S_R) \quad = \quad \int_{E_L}^{E_R} \int_{max(e,S_L)}^{S_R} h(e) \cdot f_\theta(s-e) ds de \tag{1}$$

where f_θ is the probability density function (pdf) for the incubation period and $h(e)$ is the pdf of the exposure time. We define θ as the parameter(s) defining the incubation period distribution and assume $h(e)$ is uniform. Using this likelihood, we can fit parametric models to the data.

A log-normal parametric regression model was fit to the data using maximum likelihood techniques based on the likelihood in equation (1) above. A log-normal distribution is commonly defined by two parameters – the median and the dispersion. Approximately two-thirds of cases develop symptoms between median/dispersion and median x dispersion. Previous research has shown that log-normal distributions fit incubation period data well [2, 3, 5]. We additionally fit Weibull and log logistic models to our data and compared these to the log normal distribution using the maximum log likelihood value. The log normal distribution was more supported by the single interval-censored version of our data than the other two distributions considered. The log normal model was also compared with a non-parametric fit to the data and the two models were largely consistent with each other. To test our assumption about last possible time of exposure for those cases who did not have a clear one, we conducted a brief sensitivity analysis, shortening the possible exposure window by subtracting as much as 1.5 days off of the existing E_R. The estimates of the median and 95^{th} percentile did not change by more than 5 hours for any of the scenarios considered.

The final analysis included 134 of 139 total observations from the outbreak. One individual was tested for 2009 H1N1 because a household member was positive and the test was positive. However this individual never became symptomatic. Four other individuals had no ascertainable time of possible exposure. Standard errors of the parameters were obtained by inverting the numerical approximation to the Hessian matrix at the maximum value on the two-dimensional likelihood surface. The delta method was used to obtain confidence intervals for the reported percentiles of the distribution.

The doubly interval-censored analyses were conducted using hand-coded programs in the R statistical software package [4]. Details on the performance of doubly interval-censored methods, especially in comparison with single interval-censored methods has been discussed at length elsewhere [5].

The data used in this analysis is presented in Table 1.

2 Generation time

The generation time (the time between successive symptom onsets in a chain of transmission) for 2009 H1N1 was estimated using data from infector/infectee pairs where a single likely infector could be identified and the date of symptom onset for both the infector and infectee was known. For each pair, the longest and shortest generation time was calculated. For example, assume that case A developed symptoms on day 1 and case B developed symptoms on day 3. Then the shortest possible generation time is one day: case A develops

symptoms at the very end of day 1 and case B develops symptoms at the very beginning of day 3. And the longest possible generation time is three days: case A develops symptoms at the beginning of day 1 and case B develops symptoms at the end of day 3. So the single interval-censored observation for this pair would be a generation time of 1 to 3 days.

We observed 16 pairs of cases that met the criteria for inclusion in the analysis. Again, we fit log normal, Weibull and log logistic parametric models to the data and the Weibull model achieved the largest likelihood of the three, given the data. The Weibull model was subsequently fit to the data using the `survreg()` function in the survival pacakge of R. The resulting parameters for the Weibull model were shape $= 2.36$ and scale $= 3.18$. The delta method was used to find the standard errors for the percentiles of the distribution.

The data used in this analysis is presented in Table 2.

3 Length of illness

The data used in this analysis is presented in Table 3.

4 Transmission parameters

4.1 Within-School Reproductive Number (R)

The within-school reproductive number for ILI among students was calculated using two methods. One based on the final size of the outbreak, and one based on the rate of exponential growth.

4.1.1 Estimates from the Percentage of Students Reporting ILI

We estimated R_0 from the final size of the epidemic (as measured by the percentage of students reporting ILI on the online survey using the formula presented in Vynnycky 2007 [6], and assuming a completely susceptible population:

$$R_0 = \frac{N-1}{C} \sum_{i=N-C+1}^{N-1} \frac{1}{i}$$

where $N = 2,225$ is the population at risk (number of students responding to the survey), and $C = 780$ is the number of cases (students reporting ILI). Variance was calculated as presented in [6].

4.1.2 Exponential Growth

We estimated the within-school reproductive number based on the exponential growth rate in incident cases among students, r [7, 8]. If the distribution of the generation time follow a gamma distribution then the reproductive number, R, can be calculated using the estimator [7]:

$$R = (1 + r\mu\nu^2)^{\frac{1}{\nu^2}}$$

where μ is the mean generation time and ν is the coefficient of variation. We fit a gamma distribution to our data and compared it with our preferred distributional fit, a Weibull distribution, to verify that they had the

same mean (2.8 days), and nearly identical coefficients of variation (0.64 versus 0.57). We used the value for ν of 0.64 for our analysis.

Estimation of r was performed assuming a pure birth process, hence the likelihood for r is [7]:

$$L(r|\boldsymbol{C}) \propto \exp\left(-r\sum_{i=0}^{t-1} C(i)\right)(1-\exp(-r))^{C(t)-C(0)}$$

As are primary period of analysis we considered the period between April 18th and April 24th 2009. As a sensitivity analysis we also considered periods of April 18-23, 18-22, 19-24, 19-23 and 19-22. The results of the analysis under each of these assumed periods is shown below.

from	to	estimate	Supported Interval
18	24	3.3	3.0, 3.6
18	23	5.2	4.5, 6.0
18	22	2.6	2.9, 4.6
19	24	3.3	2.9, 3.6
19	23	5.3	4.5, 6.2
19	22	3.6	2.9, 4.6

Supported intervals were calculated based on the range of values (r_{low}, r_{high}) such that $\forall r_i \in (r_{low}, r_{high}); \frac{r_i}{\hat{r}} > 1/32$ where $\hat{r}$ is the maximum likelihood estimate for r. This interval is similar to a frequentest 95% confidence interval.

4.2 Household transmission probability

We estimated the probability of one infectious individual infecting a susceptible household member using a Reed-Frost chain binomial model. This model assumes that the transmission of disease between an infectious and susceptible individual occurs independent of any other transmission in the household during a particular generation. We assume that transmission progresses in the household in discrete time steps or generations. We label the probability of transmission from one infectious individual to a susceptible individual during a given generation as p and the probability of escaping transmission from one infectious individual as $q = 1 - p$ (following Becker, 1989 [9]). We consider only households in which at least one household member has been infected. For each generation within a given household, the probability that the i infectious individuals infect exactly x of the s susceptible individuals is a standard binomial formula:

$$pr(I_{t+1} = x|S_t = s, I_t = i) = \frac{s!}{x!(s-x)!}p^x q^{s-x} \tag{2}$$

where S_t is the number of susceptibles exposed to I_t infectious individuals during generation t.

The probability of the epidemic chain within a household where I_t takes values $i_0 \to i_1 \to \cdots \to i_r$, at each generation up until generation r, the final generation of transmission, is given by

$$pr(i_0 \to i_1 \to \cdots \to i_r) = \frac{s_0!}{i_1!i_2!\cdots i_r!s_r!}\prod_{t=0}^{r} p^{i_t}q^{s_t}. \tag{3}$$

The Reed-Frost model assumes that the probability of escaping infection when exposed to two infectious individuals simultaneously is the same as escaping infection due to exposure to a single individual during two

generations. We can write an expression for the probability of observing outbreaks of size h in households of size n, $\theta_h(n)$ by writing down the possible chains of transmission that can lead to each size and adding them together. Assume that we have observed c_0 houses with zero cases in addition to the first case, c_1 houses with one case in addition to the first case, etc... We can write the likelihood observing a distribution of $c_0, c_1, c_2, \ldots$ among a group of households of size n as

$$l(\theta_0, \theta_1, ...\theta_n) = [\theta_0(n)]^{c_0} \cdot [\theta_1(n)]^{c_1} \cdot [\theta_2(n)]^{c_2} \cdots [\theta_n(n)]^{c_n} \tag{4}$$

We found the parameter p (and, consequently, q) that maximized the product of likelihoods across household sizes 2 through 5 using data on the total number of individuals reporting influenza like illness by survey among those households that had at least one laboratory confirmed 2009 H1N1 case. Maximization of the likelihood was done using the optim() function in R. A detailed account of this methodology can be found in [9]. The data used to estimate the transmission probability, the distribution of households by size and the number of individuals reporting symptoms is shown below.

		Number in Household						
		2	3	4	5	6	7	8
Number Infected	1	6	16	26	17	4	2	1
	2	1	4	19	5	6	0	1
	3		1	5	4	1	1	0
	4			0	3	1	0	0
	5				0	0	0	0
	6					0	0	0
	7						0	0
	8							0

5 NYC Department of Health and Mental Hygiene Staff Participating in Response or Commenting on Manuscript

Thank you to all of the DOHMH staff who aided in the outbreak response or provided comments on this manuscript:

Asha Abdool, Elizabeth Alt, Gary Beaudry, Susan Blank, James Durrah, Joseph Egger, Michael Johns, Stephen Lee, Alys Adamski, Gail Adman, Elisabeth Agbor-Tabi, Chris Aston, Peter Backman, Jennifer Baumgartner, Geri Bell, James Betz, Barbara Butts, Liqun Cai, Alejandro Cajigal, Shadi Chamany, Dan Cimini, Jim Cone, Debra Cook, Roseann Costarella, Christiana Coyle, Cherry-Ann Da Costa, Alexandria Daniels, Wanda Davis, Berta Dawkins, Arlene DeGrasse, Suzie DeGrechie, Otto Del Cid, Bisram Deocharan, Catherine Dentinger, Luis Diaz, Kathleen DiCaprio, Laura DiGrande, Zadkijah Edghill, Barbara Edwin, Martin Evans, Shannon Farley, Amber Featherstone, Richard Feliciano, Marcial Fernandez, Christine Fils-Aime, Ana-Marie Fireteanu, Kelly Fitzgerald, Thomas Frieden, Lawrence Fung, Stephen Giannotti, Latchmidat Girdharrie, Michelle Glaser, Chris Goranson, Leena Gupta, Bruce Gutelius, Carol Hamilton, Tiffany Harris, Ian Hartman-O'Connell, Qazi Hasnain, Michael Heller, Debra Hendrickson, Arnold Herskovitz, Kinjia Hinterland, Roosevelt Holmes, Jeanne Hom, Jeffrey Hon, Tana Hopke, Scott Hughes, Stephen Immerwahr, Anne Marie Incalicchio,

Charu Jain, John Jasek, Julia Jimnez, Nimi Kadar, Chrispin Kambili, Jisuk Kang, Deborah Kapell, Adam Karpati, Ram Koppaka, John Kornblum, Ellen Lee, Lillian Lee, Gilbert Marin, Thomas Matte, Maushumi Mavinkurve, Rene McAnanama, Ryan McKay, Carolyn McKay, Eirc Medina, Wanda Medina, Danielle Michelangelo, Juliet Milhofer, Irina Milyavskaya, Mark Misener, Joseph Mizrahi, Linda Moskin, Maria Mosquera, Matt Motherwell, Christa Myers, Yuk-Wah (Fran) Ng, Diana Nilsen, Beth Nivin, William Olesko, Sheila Palevsky, Lucille Palumbo, Peter Papadopoulos, Hilary Parton, Jacob Paternostro, Lynn Paynter, Krystal Perkins, Sharon Perlman, Haresh Persaud, Charles Peters, Vickie Peters, Lindsay Pool, Amado Punsalang, Zahedur Rasul, Valerie Rawlins, Adam Readhead, Anne Rinchiuso, Teresa Rodrguez, Ramon Rosal, Maureen Ryan, Michael Sanderson, Allison Scaccia, Amber Levanon Seligson, Jantee Seupersad, Joanne Severe-Dildy, Asma Siddiqi, Tejinder Singh, Crystal Simmons, Meredith Slopen, Timothy Snuggs, David Starr, Catherine Stayton, Jacqueline Terlonge, Catherine Travers, Kimberly Turner, Joan Tzou, Elaine Vernetti, Shameeka Vines, Donald Walker, Stanley Wang, Connie Warner, Antoinette Williams-Akita, Elisha Wilson, Eliza Wilson, Marie Wong, David Yang, Mohammad Younis, Sulaimon Yusuff.

References

[1] Carrat F, Vergu E, Ferguson NM, Lemaitre M, Cauchemez S, Leach S, Valleron AJ. Time Lines of Infection and Disease in Human Inuenza: A Review of Volunteer Challenge Studies. *Am J Epidemiol* 2008; 167:775–785.

[2] Sartwell PE. The distribution of incubation periods of infectious disease. *Am J Hyg* 1950; 51:310–8.

[3] Cowling BJ, Muller MP, Wong IOL, Ho L, Louie M, McGeer A, Leung G. Alternative methods of estimating an incubation distribution: examples from severe acute respiratory syndrome. *Epidemiology* 2007; 18(2):253–9.

[4] Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat* 1996; 5(3):299–314.

[5] Reich NG, Lessler J, Cummings DAT, Brookmeyer R. Estimating incubation period distributions with coarse data. *Stat Med* 2009; 28(22):Pages 2769–2784.

[6] Vynnycky E, Tindall A, Punam M. Estimating the reproduction numbers of Spanish influenza using morbidity data. *International Journal of Epidemiology* 2007; 36:881-889.

[7] Nishiura H, Castillo-Chavez C, Safan M, Chowell G. Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan. *Euro Surveill* 2009; 14(22).

[8] Wallinga J, Lispsitch M. How generaiton intervals shape the relationship beween growth rates and reproductive rates. *Proc R Soc B* 2007; 274:599-604.

[9] Becker, N.G., Analysis of infectious disease data, 1989, Chapman and Hall, Boca Raton.

6 Data

Table 1: Incubation period data. E_L and E_R are the times that bound the interval of possible infection times. S_L and S_R are the times that bound the interval of possible symptom onset times. All are relative to a common "time 0" in calendar time.

id	E_L	E_R	S_L	S_R
1	0	10	9	10
2	7.25	11	10	11
3	7.25	11	10	11
4	7.25	11	10	11
5	7.25	9	8	9
6	7.25	10	9	10
7	7.25	11	10	11
8	7.25	11	10	11
9	9	12	11	12
10	7.25	10	9	10
11	8	15	14	15
12	7.25	11	10	11
13	9	14	13	14
14	7.25	11	10	11
16	7.25	11	10	11
17	9	15	14	15
18	9	12	11	12
19	9	15	14	15
20	7.25	11	10	11
21	7.25	12	11	12
22	7.25	11	10	11
23	7.25	11	10	11
24	7.25	13	12	13
25	11	12	13	14
26	10	11	13	14
27	11	12	13	14
28	7.25	13	12	13
29	10	12	11	12
30	11	14	13	14
31	10	16	15	16
32	12	13	14	15
33	7.25	11	10	11
35	7.25	12	11	12
36	7.25	10	9	10
37	7.25	11	10	11
38	7.25	11	10	11
39	7.25	14	13	14
40	7.25	11	10	11
41	7.25	14	13	14
42	7.25	12	11	12
43	7.25	11	10	11
44	7.25	11	10	11
45	7.25	10	9	10
46	7.25	10	9	10
47	7.25	10	9	10
48	7.25	9	8	9
49	7.25	10	9	10
50	7.25	12	11	12
51	7.25	11	10	11
52	7.25	9	10	11
53	7.25	10	9	10
54	7.25	11	10	11
55	7.25	13	12	13
56	7.25	11	10	11
57	7.25	11	10	11
58	7.25	11	10	11
59	7.25	12	11	12
60	7.25	11	10	11
61	7.25	11	10	11
62	7.25	10	9	10
63	7.25	14	13	14
64	7.25	12	11	12
65	7.25	10	9	10
66	7.25	11	10	11
67	7.25	19	18	19
68	7.25	11	10	11
69	7.25	11	10	11
70	7.25	11	10	11
71	7.25	13	12	13

id	E_L	E_R	S_L	S_R
72	7.25	11	10	11
73	7.25	11	10	11
74	7.25	14	13	14
75	7.25	11	10	11
76	7.25	11	10	11
77	7.25	12	11	12
78	7.25	12	11	12
79	7.25	14	13	14
80	7.25	11	10	11
81	7.25	13	12	13
82	7.25	13	12	13
83	7.25	12	11	12
84	7.25	11	10	11
85	7.25	13	12	13
86	7.25	10	9	10
87	7.25	11	10	11
88	7.25	12	11	12
89	7.25	12	11	12
90	7.25	10	9	10
91	7.25	14	13	14
92	7.25	11	10	11
93	7.25	15	14	15
94	7.25	18	17	18
95	7.25	13	12	13
96	7.25	11	10	11
97	7.25	14	13	14
98	7.25	13	12	13
99	7.25	10	9	10
100	7.25	13	12	13
101	7.25	13	12	13
102	7.25	10	9	10
103	7.25	10	9	10
104	7.25	11	10	11
105	7.25	11	10	11
106	7.25	13	12	13
107	7.25	12	11	12
108	7.25	17	16	17
109	7.25	11	10	11
110	7.25	11	10	11
111	7.25	11	10	11
112	7.25	11	10	11
113	7.25	11	10	11
114	7.25	10	9	10
115	7.25	11	10	11
116	7.25	12	11	12
117	7.25	12	11	12
118	7.25	12	11	12
119	7.25	11	10	11
120	7.25	11	10	11
121	7.25	11	10	11
122	7.25	10	9	10
123	7.25	12	11	12
124	7.25	11	10	11
125	7.25	11	10	11
126	7.25	13	12	13
127	7.25	12	11	12
128	7.25	11	10	11
129	7.25	11	10	11
130	7.25	12	11	12
131	7.25	12	11	12
132	7.25	11	10	11
134	7.25	11	10	11
135	7.25	10	9	10
136	7.25	10	9	10
137	7.25	12	11	12

Table 2: Generation time data. GT_L and GT_R represent the shortest and longest possible generation times for a given pair of cases.

id	GT_L	GT_R
2	1	3
3	1	3
4	1	3
7	0	2
9	0	2
11	4	6
13	2	4
17	3	5
18	0	2
19	3	5
21	0	2
25	3	5
26	4	6
27	1	3
31	3	5
32	3	5

Table 3: Symptom duration data. D_L and D_R represent the shortest and longest possible symptom duration for a given case. When no upper bound on the duration was observed, the upper limit is indicated by ∞.

id	D_L	D_R
1	0	2
2	0	1
3	0	1
4	3	∞
5	4	∞
6	0	6
7	1	3
8	3	∞
9	5	∞
10	4	∞
11	2	∞
12	3	∞
13	3	∞
14	0	16
16	4	∞
17	2	∞
18	5	∞
19	0	12
20	3	∞
21	13	∞
22	7	∞
23	3	∞
24	4	6
25	9	∞
26	9	11
27	2	∞
28	2	4
29	11	∞
30	3	5
31	2	4
32	3	5
33	5	∞
34	1	∞
35	7	9
36	4	∞
37	3	∞
38	7	∞
40	3	5
41	0	2
42	5	∞
43	3	∞
44	5	∞
45	4	∞
46	4	∞
47	4	∞
48	5	∞
49	4	∞
50	3	5
51	3	∞
52	3	∞
53	4	∞
54	3	∞
55	5	7
56	3	∞
57	7	9
58	7	∞
59	0	15
60	3	5
61	4	6
62	4	∞
63	0	13
64	16	∞
65	6	8
66	3	∞
68	2	4
69	3	5
70	7	9
71	0	17
72	6	∞
73	0	5

id	D_L	D_R
74	0	13
75	4	6
76	0	16
77	4	∞
78	3	5
79	0	16
81	0	7
82	0	7
83	2	∞
84	5	7
85	3	5
86	5	7
87	18	∞
88	7	∞
89	0	20
90	4	6
91	0	16
92	3	∞
93	0	12
94	6	8
95	5	7
96	5	7
97	11	∞
98	2	4
99	0	17
100	0	17
101	0	7
102	6	8
103	4	∞
104	0	5
105	7	∞
106	3	5
107	2	4
108	1	3
109	3	∞
110	0	16
111	3	∞
112	3	∞

id	D_L	D_R
113	3	∞
114	4	∞
115	10	∞
116	3	5
117	6	∞
118	2	∞
119	3	∞
120	0	19
121	3	5
122	11	13
123	1	3
124	3	∞
125	3	∞
126	4	6
127	1	3
128	3	∞
129	3	∞
130	3	5
131	4	∞
132	3	∞
133	8	10
134	3	∞