Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings

Christophe Fraser,1,* Christi A. Donnelly,1 Simon Cauchemez,1 William P. Hanage,1 Maria D. Van Kerkhove,2 T. Deirdre Hollingsworth,3 Jamie Griffin,1 Rebecca F. Baggaley,1 Helen E. Jenkins,1 Emily J. Lyons,7 Thibaut Jombart,3 Wes R. Hinsley,7 Nicholas C. Grassly,1 Francois Baloux,1 Azra C. Ghani,5 Neil M. Ferguson1†;
Andrew Rambaut,2 Oliver G. Pybus3;
Hugo Lopez-Gatell,4 Celia M. Alpuche-Araujo,5 Jetzia Bojorquez Chapela,4 Ethel Palacios Zavala4;
Dulce Ma. Espejo Guevara6;
Francesco Checchi,1 Erika Garcia,3 Stephane Hugonnet,7 Cathy Roth7

The WHO Rapid Pandemic Assessment Collaboration†:

A novel influenza A (H1N1) virus has spread rapidly across the globe. Judging its pandemic potential is difficult with limited data, but nevertheless essential to inform appropriate health responses. By analyzing the outbreak in Mexico, early data on international spread, and viral genetic diversity, we make an early assessment of transmissibility and severity. Our estimates suggest that 23,000 (range 6000 to 32,000) individuals had been infected in Mexico by late April, giving an estimated case fatality ratio (CFR) of 0.4% (range: 0.3 to 1.8%) based on confirmed and suspected deaths reported to that time. In a community outbreak in the small community of La Gloria, Veracruz, no deaths were attributed to infection, giving an upper 95% bound on CFR of 0.6%. Thus, although substantial uncertainty remains, clinical severity appears less than that seen in the 1918 influenza pandemic but comparable with that seen in the 1957 pandemic. Clinical attack rates in children in La Gloria were twice that in adults (<15 years of age: 61%; ≥15 years: 30%). Three different epidemiological analyses gave basic reproduction number (R0) estimates in the range of 1.4 to 1.6, whereas a genetic analysis gave a central estimate of 1.2. This range of values is consistent with 14 to 73 generations of human-to-human transmission having occurred in Mexico to late April. Transmissibility is therefore substantially higher than that of seasonal flu, and comparable with lower estimates of R0 obtained from previous influenza pandemics.

On 29 April 2009, the World Health Organization (WHO) announced that the rapid global spread of a strain of influenza A (H1N1) virus detected in the previous week warranted moving the global pandemic alert level to phase 5 (www.who.int/csr/disease/swineflu/). Phase 5 indicates sustained human-to-human transmission of a novel influenza strain of animal origin in one WHO region of the world, and exported cases detected in other regions. In this outbreak, the earliest affected country may have been Mexico, with many cases in other nations associated with these.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: neil.ferguson@imperial.ac.uk
‡All authors are members of this collaboration.

REPORTS
and assuming similar times from infection to confirmation and from infection to death, we estimated CFRs in the range of 0.3 to 0.6% from the interval-censored case count model, based on confirmed and suspected deaths combined, or 0.03 to 0.05% for confirmed deaths only. Using the alternative, more pessimistic, country presence/absence model, we estimated CFRs of 0.9 to 1.8% based on suspected and confirmed deaths, and 0.08 to 0.16% from the confirmed deaths alone. These estimates have already changed somewhat as a result of data available after 30 April, but we deliberately report the earlier analysis because it formed part of the evidence base used by WHO to move to phase 5.

Another source of information on severity comes from the large outbreak of respiratory disease seen in the small, isolated community of La Gloria in Veracruz province, one case of which has been confirmed to have been caused by the novel H1N1 strain. It is possible that other viruses were circulating at the same time as the outbreak, but the overall attack rate is substantially larger than would be expected for a seasonal influenza outbreak. No fatalities among 616 cases have been attributed to infection during the full period of surveillance of that outbreak (Fig. 3A), giving a 95% confidence interval (CI) of 0 to 0.6%.

Data on the magnitude of the current outbreak in Mexico can also be used to estimate the transmissibility of the virus if the start date of the outbreak is known or can be estimated. Epidemiological investigations into the emergence of the virus in Mexico have focused on the La Gloria outbreak, where the first case in that outbreak is thought to have occurred around 15 February 2009 (Fig. 3A).

An alternative approach to estimating the start date of the outbreak is to look at the diversity in the genetic sequences of viral samples collected from confirmed cases, assuming that diversity accumulates according to a molecular clock model. Twenty-three complete publicly available hemagglutinin (HA) gene sequences from cases not linked in epidemiological clusters were analyzed with a Bayesian coalescent method that assumes exponential growth of the viral population (7). This yielded an estimate of the time of most recent common ancestor (TMRCA) of 12 January 2009 [95% credible interval (CrI): 3 November 2008 to 2 March 2009]. The genetic model also gave an estimate of the doubling time of the epidemic of 10 days (95% CrI: 4.5 to 37.5 days) (Fig. 2). Assuming exponential growth, the TMRCA is a reasonable estimate of the start of the outbreak, although it is formally an upper bound due to incomplete sampling of the epidemic and the effects of the exponential model prior to distribution. These findings from a population genetic analysis are consistent with the epidemiological investigation of both the start and magnitude of the current epidemic in Mexico. Figure 2 also shows a preliminary version of this analysis based on the first 11 sequences, which gave similar estimates highlighting the power of these methods. [See (8) for further sensitivity analysis and methods.]

The reproduction number, defined as the number of cases one case generates on average over the course of their infectious period, is a key measure of transmissibility and can be estimated in a number of ways from the data currently available.

First, by assuming exponential growth, the growth rate of the epidemic (r) can be inferred from estimates of the current cumulative number of infections (Y) and estimated start date and size for the outbreak (t0 and Y0, respectively). The basic reproduction number (R0) can be estimated from the exponential growth rate if one also assumes that the generation time distribution for the new H1N1 strain is similar to that of other strains of seasonal and pandemic viruses (9, 10) [Table 1 and (8)]. Using the date of 15 February as the first case of the La Gloria outbreak (8) gives reproduction number estimates of between 1.31 and 1.42, depending on which variant of the geographical backcalculation model is used. Extending a more sophisticated Bayesian estimation method (11) that allows for stochastic variability intrinsic to epidemic dynamics and parameter uncertainty gave similar but slightly higher estimates for R0 with wider ranges: posterior median = 1.40; 95% CrI: 1.15 to 1.90 (Fig. 1C).

Second, by assuming a prior distribution on the generation time distribution informed by previous estimates of influenza, the Bayesian coa-

---

**Fig. 1.** (A). The number of passengers flying out of Mexico by actual destination and the number of confirmed cases as reported on 30 April 2009. (B) The number of cases exported to country j as reported on 30 April 2009 as a function of the estimated average number of foreign travelers in Mexico from country j on any given day in March or April. Black circles: minimal number based on one exposure per epidemiological cluster; filled red circles, total number of confirmed cases. (C) Mean assumed generation time distribution (red) and 100 illustrative draws from the prior distribution, and (D) corresponding posterior distribution of R0 estimates for a stochastic model of an epidemic within Mexico with travelers infected at a rate proportional to the estimated density of travelers per local resident. The two bar charts correspond to a 7-day delay between infection and confirmation (blue) and no delay (orange) in cases among travelers. (E) Number of acute respiratory infection cases per 100,000 inhabitants by state as reported on 5 May 2009 (1), demonstrating spatial distribution of disease within Mexico.
lescent population genetic analysis yielded a second set of estimates for $R_0$: posterior median = 1.22; 95% CrI: 1.05 to 1.60 (Fig. 2C).

Third, $R_0$ can also be estimated from analysis of the dynamics of the epidemic within defined settings. Detailed data collected by the Mexican authorities investigating the La Gloria outbreak indicate that 616 individuals from a resident population of 1575 had acute respiratory infection between 15 February and 14 April 2009 (Fig. 3A). Data on the age distribution of cases and the dates of disease onset were used in our analysis. Figure 3B shows that the clinical attack rate varied markedly as a function of age, with 61% of individuals under 15 years old affected, dropping to 29% of people over that age. The corresponding relative risk is 2.13, with a 95% CI of 1.89 to 2.39. Based on all confirmed cases in Mexico as reported on 5 May 2009 (1), the corresponding relative risk is 1.52 (95% CI: 1.33 to 1.73). The overall community attack rates seen in La Gloria are comparable to (or higher than) those seen in previous pandemics (12).

Fitting alternative epidemic models to the La Gloria data (8) demonstrated that a model with heterogeneous mixing by age plus age-dependent susceptibility to infection was required to adequately fit the data with plausible parameter estimates. The resulting maximum-likelihood estimate of $R_0$ was 1.58 with a 95% CI of 1.34 to 2.04 (Table 2). This analysis also provided the only independent estimate of the mean generation time, $T_g$ (1.91 days; 95% CI: 1.30 to 2.71 days) (Table 2), shorter than earlier estimates for influenza (9, 10), though not significantly so. It is biologically plausible that $R_0$ and $T_g$ could be correlated, because both are linked to the underlying replicative fitness of the virus. More data are needed. Owing to parameter identifiability issues, it was not possible to estimate age-dependent infectiousness, as well as age-dependent mixing, from these data. Although these estimates are informative, it should be emphasized that some uncertainties remain regarding the denominator population and that a range of other models may fit the data as well as the model choice shown here. Household data would be particularly useful in reducing remaining uncertainty.

Fourth, the time-dependent reproduction number ($R_t$) can be estimated from the time series of reported disease onsets among confirmed cases in Mexico (Fig. 4). These data are subject to much uncertainty because of marked changes in surveillance over the reporting interval, plus the nonspecificity of symptoms that are similar to existing and perhaps simultaneously circulating strains of influenza. However, we developed methods for analyzing such data (8) that account for substantial underreporting, with a change in the underreporting rate from 17 April when surveillance within Mexico was intensified. The average value of $R_t$ estimated for Mexico up until the end of April was 1.37 (95% CrI: 1.24 to 1.59) for a model with Poisson case counts, and 1.47 (95% CrI: 1.21 to 1.88) for a perhaps more plausible negative binomial model allowing daily case counts to be overdispersed (8).

Given estimates of $R_0$ and the current epidemic size $x$, we can estimate the number of generations $N_t$ of transmission of the virus among humans that is necessary to explain the current epidemic. Assuming a simple branching process with reproduction number $R_0$, the mean number of generations of transmission is given by $N_t = \ln(x/x_0)/\ln(R_0)$, assuming the epidemic was started by $x_0$ humans being infected from animal sources. Assuming $x_0 = 1$ gives estimates of $N_t$ between 14 and 73. But even if we assume that 5% of cases were infected directly from animal sources, we obtain an estimated 5 to 22 generations of transmission, indicating sustained human-to-human transmission in Mexico.

All of the $R_0$ estimates are comparable with, but perhaps on the low end of, $R_0$ estimates...
societal cost of interventions, such as school closures remains the balancing of the economic and public health benefits of mitigating the pandemic (22). Tailed computer simulations used to study scenarios consistent with the lowest values used in earlier deliberations showed the exponential growth rate of the outbreak in Mexico (Fig. 3). For sensitivity analysis and model selection, we tested several reduced model variants. None fitted significantly worse, but several produced implausible estimates of the generation time. The respective best-fit values are as follows: 1. No asymptomatics and no misreporting, no assortative mixing: $R_0 = 1.4$, $\theta = 0$, $R_f = 1.37$, $T_g = 1.39$, and $\rho_{\text{child}} = 2.80$. 2. No asymptomatics and no misreporting: $p_{\text{sym}} = 1$, $R_0 = 1.41$, $T_g = 1.53$, $\theta = 0.31$, and $\rho_{\text{child}} = 2.22$. 3. No assortative mixing: same as variant 1. 4. Model with long generation time consistent with previous estimates from influenza (also shown in Fig. 3), $T_g = 2.60$ days, $R_0 = 1.97$, $\rho_{\text{child}} = 2.52$, $\theta = 0.51$, and $p_{\text{sym}} = 0.72$. (The symbol “$\equiv$” denotes parameters defined to take fixed values.)

Table 1. Parameter estimates (and 95% confidence intervals) for the cumulative number of influenza A (H1N1) infections in Mexico among Mexican residents by late April 2009, along with corresponding estimates for the case fatality ratio (CFR), the basic reproduction number, and the exponential growth rate, assuming that infections in travelers occurred in Mexico.

<table>
<thead>
<tr>
<th>Assumed average duration of stay for visitors arriving by flights</th>
<th>Estimated number of infections among Mexican residents</th>
<th>Case fatality ratio*</th>
<th>Estimated reproduction number $R_0$‡</th>
<th>Estimated exponential growth rate$\dagger$ (doubling time in days$\dagger$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assuming 9 deaths§</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assuming 101 deaths§</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on analysis of interval-censored country counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 days</td>
<td>32,000</td>
<td>0.03%</td>
<td>0.32%</td>
<td>1.42</td>
</tr>
<tr>
<td></td>
<td>(26,000, 39,000)</td>
<td>(0.02%, 0.03%)</td>
<td>(0.26%, 0.39%)</td>
<td>(5.6)</td>
</tr>
<tr>
<td>9 days (23)</td>
<td>23,000</td>
<td>0.04%</td>
<td>0.44%</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>(20,000, 280,00)</td>
<td>(0.03%, 0.05%)</td>
<td>(0.37%, 0.52%)</td>
<td>(5.9)</td>
</tr>
<tr>
<td>12 days</td>
<td>180,00</td>
<td>0.05%</td>
<td>0.55%</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>(150,000, 220,00)</td>
<td>(0.04%, 0.06%)</td>
<td>(0.46%, 0.66%)</td>
<td>(6.1)</td>
</tr>
<tr>
<td>Based on analysis of presence or absence of confirmed disease in countries</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 days</td>
<td>11,000</td>
<td>0.08%</td>
<td>0.90%</td>
<td>1.36</td>
</tr>
<tr>
<td></td>
<td>(5,000, 27,000)</td>
<td>(0.03%, 0.20%)</td>
<td>(0.37%, 2.20%)</td>
<td>(6.6)</td>
</tr>
<tr>
<td>9 days (23)</td>
<td>7,000</td>
<td>0.12%</td>
<td>1.36%</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>(3,000, 18,000)</td>
<td>(0.05%, 0.29%)</td>
<td>(0.56%, 3.30%)</td>
<td>(7.0)</td>
</tr>
<tr>
<td>12 days</td>
<td>6,000</td>
<td>0.16%</td>
<td>1.81%</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>(2,000, 14,000)</td>
<td>(0.07%, 0.39%)</td>
<td>(0.74%, 4.40%)</td>
<td>(7.4)</td>
</tr>
</tbody>
</table>

*This is a simple estimate of the CFR (6). The numbers of deaths and the data on cases in countries were both obtained as reported on 30 April. Although censoring is important to consider when estimating the CFR (due to the time interval between case report and death), in these data we have both the time between clinical onset and death within Mexico and the time between clinical onset and case confirmation in other countries to consider. Given uncertainty surrounding both of these distributions, we have made the simplifying assumption that these time intervals are similar and thus that the CFR can be estimated by the deaths within Mexico attributed to the novel influenza A (H1N1) divided by the estimated number of infections among Mexican residents based on data reported up to 30 April. §Based on 9 confirmed and 92 suspected deaths (101 total) that were reported by 30 April 2009 (6). ‡These estimates assume that the estimated cumulative number of influenza A (H1N1) infections in Mexico among Mexican residents relates to the cumulative number of infections up to 23 April 2009 (i.e., 1 week before the data were reported).

Table 2. Epidemiological parameters estimated by fitting an age-stratified mathematical model to the outbreak in the village of La Gloria (Fig. 3). For sensitivity analysis and model selection, we tested several reduced model variants. None fitted significantly worse, but several produced implausible estimates of the generation time. The respective best-fit values are as follows: 1. No asymptomatics and no misreporting, no assortative mixing: $R_0 = 1.4$, $\theta = 0$, $R_f = 1.37$, $T_g = 1.39$, and $\rho_{\text{child}} = 2.80$. 2. No asymptomatics and no misreporting: $p_{\text{sym}} = 1$, $R_0 = 1.41$, $T_g = 1.53$, $\theta = 0.31$, and $\rho_{\text{child}} = 2.22$. 3. No assortative mixing: same as variant 1. 4. Model with long generation time consistent with previous estimates from influenza (also shown in Fig. 3), $T_g = 2.60$ days, $R_0 = 1.97$, $\rho_{\text{child}} = 2.52$, $\theta = 0.51$, and $p_{\text{sym}} = 0.72$. (The symbol “$\equiv$” denotes parameters defined to take fixed values.)

<table>
<thead>
<tr>
<th>Best estimate</th>
<th>95% confidence interval</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_f$</td>
<td>1.58</td>
<td>Basic reproduction number</td>
</tr>
<tr>
<td>$T_g$</td>
<td>1.91</td>
<td>Mean generation time (days)</td>
</tr>
<tr>
<td>$p_{\text{sym}}$</td>
<td>86%</td>
<td>Proportion of cases that are symptomatic and ascertained</td>
</tr>
<tr>
<td>$\rho_{\text{child}}$</td>
<td>2.06</td>
<td>Susceptibility of children relative to adults</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.50</td>
<td>Assortativity of mixing between children and adults (0 = random, 1 = fully assortative)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assumed value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_c$</td>
<td>Fraction of the generation time that is latent (uninfected)</td>
</tr>
<tr>
<td>$\phi_{\text{child}}$</td>
<td>Assumed Infectiousness of children relative to adults</td>
</tr>
</tbody>
</table>

obtained from analysis of previous pandemics [1.4 to 2.0 for 1918, 1957, and 1968 (9, 13–15)].

Overall, our transmissibility estimates are consistent with the lowest values used in earlier detailed computer simulations used to study scenarios in pandemic mitigation (16, 17), indicating that the conclusions regarding control policy effectiveness reached by those analyses could be relevant to the current epidemic. However, the key trade-off remains the balancing of the economic and societal cost of interventions, such as school closure, against the numbers of lives saved through use of such measures. Where substantial antiviral stockpiles are available, a secondary trade-off is the extent to which large-scale prophylaxis is justified, given the potential risks of high-level resistance developing (18–21). At present, estimates of disease severity are insufficiently robust to allow these trade-offs to be properly evaluated, but that uncertainty should diminish rapidly in coming weeks as more data on severe cases in the United States and other countries become available.

As the situation develops, a key issue is to optimize study designs and surveillance protocols to be most informative in estimating some of these unknown factors, thus potentially informing and refining the public health response. Clearly, detailed investigations of transmission in households and schools will be useful, as would be the consistent collection and dissemination of electronic patient records, which could be used to detect cofactors in the severity of infection.

In conclusion, while the emerging data from Mexico and other countries have enabled important insights into the origin, extent, transmissibility, and severity of the unfolding pandemic [including detailed epidemiological analysis of data from the U.S. outbreak recently published (22)], many uncertainties remain and should not be underestimated. The incubation and infectious periods have not yet been reliably ascertained, leaving uncertainty in estimates of the generation time. Much remains to be done to estimate clinical severity of infection, to understand regional variations seen so far (or indeed, whether they exist). As the epidemic spreads further, it is likely that severity will vary from country to country depending on health care resources and the public health measures adopted to mitigate impact. The existence of any cross-immunity (perhaps not mediated via HA-specific antibodies) from past exposure to prior influenza A subtypes is unknown, but the strong age dependence in clinical attack rates seen in La Gloria is intriguing. Cross-immunity would imply that $R_f$ could be higher in fully susceptible populations than estimated here.
The future evolution of the transmissibility, antigenicity, virulence, and antiviral resistance profile of this or any influenza virus is difficult to predict. It is also unclear whether this strain will displace existing influenza A subtypes from the human population, as occurred in the past three pandemics. The extent to which seasonal damping of transmission in North America and Europe is responsible for the moderate transmissibility seen to date is uncertain; the progress of transmission in the Southern Hemisphere (which is just entering its influenza season) needs to be carefully monitored in the next few months. To reduce all these uncertainties, it is essential that public health agencies around the world continue to collect high-quality epidemiological data in a focused, resource-efficient manner despite the expected increases in case numbers in coming weeks. Epidemiological analysis and modeling are useful tools for guiding such efforts and interpreting the resulting data.

Note added in proof: We cited two sources (1, 6) for confirmed and suspected deaths in Mexico, reported by 4 May 2009 and 30 April 2009, respectively. These sources are not publicly available at present. However, similar reports are publicly available: The Mexican government Web site (24) gives some data on the 5 May situation report (25) documenting 26 confirmed deaths and 114 suspected deaths (77 without samples for analysis), and Mortality and Morbidity Weekly Report (26) lists 7 confirmed and 77 suspected deaths posted on 30 April. Since this article appeared online, the number of deaths in Mexico up to 23 April has been determined to be 21, resulting in a revised estimate of the CFR of 0.091% (range: 0.066 to 0.35%) (24).

References and Notes
8. Additional details on methods, data, and results are in the Supporting Online Material.
23. L. E. Hudman, R. H. Jackson, Geography of Travel and Tourism (Engage Learning, Thomson Learning, Clifton Park, NY, ed. 2002).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1176062/DC1
Methods
Figs. S1 to S3
Tables S1 to S12
References
Epidemiological data
5 May 2009; accepted 11 May 2009
Published online 11 May 2009
10.1126/science.1176062
Include this information when citing this paper.

Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees

Kevin Liu,1 Sindhu Raghavan,2 Serita Nelesen,1 C. Randal Linder,1 Tandy Warnow1*

Inferring an accurate evolutionary tree of life requires high-quality alignments of molecular sequence data sets from large numbers of species. However, this task is often difficult, slow, and idiosyncratic, especially when the sequences are highly diverged or include high rates of insertions and deletions (collectively known as indels). We present SATé (simultaneous alignment and tree estimation), an automated method to quickly and accurately estimate both DNA alignments and trees with the maximum likelihood criterion. In our study, it improved tree and alignment accuracy compared to the best two-phase methods currently available for data sets of up to 1000 sequences, showing that coestimation can be both rapid and accurate in phylogenetic studies.

Phylogeny estimation from molecular sequences typically has two phases: An alignment is estimated, and then a tree is produced for the alignment. Alignment methods like MAFFT (1), Probcons (2), Probtree (3), Prank (4), and Muscle (5) provide more accurate alignments than earlier methods (3, 4, 6), and maximum likelihood (ML) methods of phylogeny estimation [e.g., RAxML (7, 8), GARLI (9), and Phyml (10)] produce more accurate trees for large data sets than other