Inference of seasonal and pandemic influenza transmission dynamics

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1073/pnas.1415012112</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:26882153">http://nrs.harvard.edu/urn-3:HUL.InstRepos:26882153</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Inference of seasonal and pandemic influenza transmission dynamics

Wan Yang\textsuperscript{a,1}, Marc Lipsitch\textsuperscript{b}, and Jeffrey Shaman\textsuperscript{a}

\textsuperscript{a}Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY 10032; and \textsuperscript{b}Center for Communicable Disease Dynamics, Harvard School of Public Health, Harvard University, Boston, MA 02115

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved January 22, 2015 (received for review August 5, 2014)

The inference of key infectious disease epidemiological parameters is critical for characterizing disease spread and devising prevention and containment measures. The recent emergence of surveillance records mined from big data such as health-related online queries and social media, as well as model inference methods, permits the development of new methodologies for more comprehensive estimation of these parameters. We use such data in conjunction with Bayesian inference methods to study the transmission dynamics of influenza. We simultaneously estimate key epidemiological parameters, including population susceptibility, the basic reproductive number, attack rate, and infectious period, for 115 cities during the 2003–2004 through 2012–2013 seasons, including the 2009 pandemic. These estimates discriminate key differences in the epidemiological characteristics of these outbreaks across 10 y, as well as spatial variations of influenza transmission dynamics among subpopulations in the United States. In addition, the inference methods appear to compensate for observational biases and underreporting inherent in the surveillance data.

Influenza remains a significant public health burden. Worldwide, it causes 3–5 million cases of severe illness, and 250–500 thousand deaths annually (1). Several times per century, a novel antigenic subtype appears and is able to spread efficiently from person to person, causing a pandemic. Influenza epidemics recur annually because influenza A viruses, especially, continually evolve to escape human immunity, leaving a proportion of the population newly susceptible despite prior infections. These epidemics occur generally in the winter in temperate regions and show more temporal variability in the tropics. In many countries, traditional sentinel surveillance systems track the incidence of influenza-like illness (ILI) and the prevalence of PCR-confirmed influenza among specimens sent for testing. The product of these quantities has been proposed as a proxy for the incidence of influenza, up to an unknown multiplicative constant, under certain assumptions (2).

The Centers for Disease Control and Prevention (CDC) has collected weekly ILI surveillance data routinely for over 3 decades. Additionally, algorithms have been developed to estimate ILI based on online search queries (3); these Google Flu Trends (GFT) ILI data are available for 29 countries around the world in real time. While surveillance data are valuable for assessing influenza activity in the general population, such measures typically do not depict true influenza infection rates. For instance, ILI is recorded as the number of patients diagnosed with ILI per patient–doctor visit in the United States rather than per capita incidence. Further, underreporting is common due to asymptomatic infections and those symptomatic but unattended. These deficiencies, together with observational error, create challenges for delineating a complete picture of influenza epidemiology using these data; however, Bayesian inference methods exist that, when coupled with dynamical models, are equipped to handle such imperfect observational data and partially observed dynamical systems (4–8).

Our previous work applied such Bayesian inference frameworks (also known as data assimilation methods) to simulate seasonal outbreaks and demonstrated that reliable forecast of the peak timing of seasonal influenza can be made using GFT ILI data combined with regional viral isolation data (termed ILI+) (4, 5). Here, we use these same inference methods to estimate key epidemiological parameters for both pandemic and seasonal outbreaks, using the entirety of outbreaks in 115 US cities during the 2003–2004 through 2012–2013 influenza seasons, and demonstrate how big-data-driven surveillance can be used to reveal the transmission dynamics of influenza among the general population.

Results

Several epidemiological parameters are key for understanding the transmission dynamics of an infectious disease. The basic reproductive number, $R_0$, represents the average number of secondary infections generated from a primary case in an entirely susceptible population. Characterization of $R_0$ is consequently critical for assessing the rate of disease propagation through a population and the possibility of outbreak containment. For instance, studies have suggested that pandemic influenza spread can be contained if $R_0 < 2$ (9, 10). When analyzed retrospectively, $R_0$ facilitates quantification of the variability of influenza transmission potential among different outbreaks.

Epidemiological surveys and mathematical modeling approaches have been used to estimate $R_0$. These estimates often rely on surveillance data during the first few weeks of an outbreak, i.e., the exponential growth period. For influenza, $R_0$ has been estimated at values between 1.3 and 3 (11), depending on the specific assumptions made with respect to initial susceptibility and either serial interval or infectious period. While some estimates of these two parameters track the incidence of influenza-like illness (ILI) and the prevalence of PCR-confirmed influenza among specimens sent for testing. The product of these quantities has been proposed as a proxy for the incidence of influenza, up to an unknown multiplicative constant, under certain assumptions (2).

The Centers for Disease Control and Prevention (CDC) has collected weekly ILI surveillance data routinely for over 3 decades. Additionally, algorithms have been developed to estimate ILI based on online search queries (3); these Google Flu Trends (GFT) ILI data are available for 29 countries around the world in real time. While surveillance data are valuable for assessing influenza activity in the general population, such measures typically do not depict true influenza infection rates. For instance, ILI is recorded as the number of patients diagnosed with ILI per patient–doctor visit in the United States rather than per capita incidence. Further, underreporting is common due to asymptomatic infections and those symptomatic but unattended. These deficiencies, together with observational error, create challenges for delineating a complete picture of infectious disease epidemiology. This issue is true for influenza, an infectious disease of pandemic potential. Here we develop and present influenza inference systems capable of compensating for observational biases and underreporting. Using both Google Flu Trends and Centers for Disease Control and Prevention data in conjunction with Bayesian model inference methods, we are able to infer the evolving epidemiological features of influenza and its impacts among the large population during 2003–2013, including the 2009 pandemic. In addition, differences among regions within the United States are identified.

Author contributions: W.Y., M.L., and J.S. designed research; W.Y. performed research; J.S. compiled data; W.Y., M.L., and J.S. analyzed data; and W.Y., M.L., and J.S. wrote the paper.

Conflict of interest statement: M.L. discloses consulting or honorarium income from the World Health Organization, Novartis, GSK, Janssen, and Pfizer; and authorship on patents held by the University of California. M.L. is a member of the Advisory Board for the World Health Organization’s (WHO) Pandemic Flu Registry (Outcome Sciences, funded in part by Roche), AHRQ, National Institute of Allergy and Infectious Diseases, and Novartis. J.S. is a member of the Advisory Board for the WHO Pandemic Flu Registry. M.L. is on the editorial board of the Journal of the American Medical Association. W.Y. is a member of the Advisory Board for the Social Science Research Network and is on the editorial board of the Journal of the American Medical Association.

Significance

Infectious disease surveillance systems are powerful tools for monitoring and understanding infectious disease dynamics; however, underreporting (due to both unreported and asymptomatic infections) and observation errors in these systems create challenges for delineating a complete picture of infectious disease epidemiology. This issue is true for influenza, an infectious disease of pandemic potential. Here we develop and present influenza inference systems capable of compensating for observational biases and underreporting. Using both Google Flu Trends and Centers for Disease Control and Prevention data in conjunction with Bayesian model inference methods, we are able to infer the evolving epidemiological features of influenza and its impacts among the large population during 2003–2013, including the 2009 pandemic. In addition, differences among regions within the United States are identified.

Author contributions: W.Y., M.L., and J.S. designed research; W.Y. performed research; J.S. compiled data; W.Y., M.L., and J.S. analyzed data; and W.Y., M.L., and J.S. wrote the paper.

Conflict of interest statement: M.L. discloses consulting or honorarium income from the World Health Organization, Novartis, GSK, Janssen, and Pfizer; and authorship on patents held by the University of California. M.L. is a member of the Advisory Board for the World Health Organization’s (WHO) Pandemic Flu Registry (Outcome Sciences, funded in part by Roche), AHRQ, National Institute of Allergy and Infectious Diseases, and Novartis. J.S. is a member of the Advisory Board for the WHO Pandemic Flu Registry. M.L. is on the editorial board of the Journal of the American Medical Association. W.Y. is a member of the Advisory Board for the Social Science Research Network and is on the editorial board of the Journal of the American Medical Association.

Significance

Infectious disease surveillance systems are powerful tools for monitoring and understanding infectious disease dynamics; however, underreporting (due to both unreported and asymptomatic infections) and observation errors in these systems create challenges for delineating a complete picture of infectious disease epidemiology. This issue is true for influenza, an infectious disease of pandemic potential. Here we develop and present influenza inference systems capable of compensating for observational biases and underreporting. Using both Google Flu Trends and Centers for Disease Control and Prevention data in conjunction with Bayesian model inference methods, we are able to infer the evolving epidemiological features of influenza and its impacts among the large population during 2003–2013, including the 2009 pandemic. In addition, differences among regions within the United States are identified.
parameters have been derived from clinical and epidemiological data (12, 13), they are usually somewhat arbitrarily assumed in modeling approaches, which adds further uncertainty to any estimation of $R_0$. Particularly, the initial susceptibility, often assumed close to 100% for a pandemic strain, is less certain for seasonal strains due to difficulty constraining factors such as rates of vaccination coverage and efficacy, which can affect the population susceptibility during a given season. As a result, $R_0$ is usually estimated for pandemic outbreaks but only rarely for seasonal outbreaks.

In this study, we applied a humidity-driven Susceptible—Infected—Recovered—Susceptible (SIRS) model jointly with either the ensemble adjustment Kalman filter (EAKF) or a particle filter (PF) to simulate influenza outbreaks for 115 US cities (4, 14). Our SIRS model includes two variables—numbers of susceptible (S) and infected (I) persons—and five parameters—immunity period ($L$), infectious period ($D$), maximum and minimum basic reproductive numbers, $R_{0\text{max}}$ and $R_{0\text{min}}$, and $\gamma$, a scaling factor used to map the SIRS model simulated incidence rate to the observation metric $MI^+$(see Materials and Methods).

In this study, the initial susceptibility for a given local outbreak is defined as the maximal estimate of susceptibility, as made during simulation of that outbreak with the model inference system; the basic reproductive number $R_0$, effective reproductive number $R_e$ (defined as $R_0S/N$, where $N = 100,000$ is the population), and infectious period $D$ are estimated at the week of maximal $R_0$. Both filter methods provide joint estimates of these epidemiological variables and parameters without specific assumptions (e.g., the level of initial susceptibility). Results with the EAKF and PF were found to be consistent (see SI Appendix, Fig. S1).

Before applying this inference framework to observed US epidemics, we tested both data assimilation methods—the EAKF and the PF—using four alternate compartment model forms: (i) a SIRS model, (ii) a Susceptible—Infected—Recovered (SIR) model, (iii) a Susceptible—Exposed—Infectious—Recovered (SEIR) model, and (iv) a SEIR-type model with two submodels for both the exposed and infectious compartments (referred to as SEIR-FR). These combined model inference frameworks were tested using model-generated “synthetic” data for which target parameter estimates were known. Our tests showed that inferences made using simpler models (SIRS and SIR) outperformed those using more complicated models (SI Appendix). In addition, we found that the single-age class SIRS model inference system is able to reliably infer the leading eigenvalue of the effective reproductive number $R_e$ from aggregated time series synthesized from multiclass compartmental models (SI Appendix). Consequently, we here present results using a single-age class SIRS model.

Our estimates of key epidemiological characteristics in the United States during 2003–2013 using this SIRS model (Table 1) are generally in line with findings from past studies (10, 11, 15–22); however, by using data from 115 cities, we are able to provide a much more comprehensive assessment of these characteristics over 10 y. In doing so, we are also able to discriminate and quantify variations in these epidemiological parameters across years including both seasonal and pandemic influenza outbreaks (Fig. 1). As would be expected, population susceptibility is highest at the beginning of the spring and fall waves of 2009 pandemic, with a mean of 76.5% and an interquartile range (IQR) of 72.9–79.6%; and is significantly higher than any of the nine epidemic seasons ($P < 2.2e−16$). Seasons with moderately severe outbreaks, i.e., the 2003–2004, 2007–2008, and 2012–2013 (23, 24), had the next highest initial susceptibility estimates. In contrast, seasons right after those profound outbreaks (e.g., 2003–2004, 2007–2008, and the pandemic) had the lowest initial susceptibilities (e.g., 2010–2011, 2004–2005, and 2008–2009) (Fig. 1C).

Past studies estimate the basic reproductive number, $R_0$, for the 2009 pandemic in the range from 1.2 to 2.3 with a median of 1.5 and tend to report larger values at the beginning of the pandemic (18). We estimate $R_0$ for the spring and fall 2009 pandemic waves as 1.63 (1.41–1.79; mean and IQR, same elsewhere unless stated otherwise). Interestingly, the 2009 pandemic strain had an $R_0$ lower than the 2003–2013 interpandemic strains (Fig. 1C). The 2003 A/H3N2-Fujian epidemic had the highest $R_0$ estimates. Our modeling framework assumes that $R_0$ varies as a function of ambient absolute humidity conditions (see Materials and Methods). This effect implies that for a given outbreak, $R_0$ will typically maximize in deep winter when absolute humidity levels are lowest; however, due to higher population susceptibility, the 2009 pandemic outbreaks took place out of season when humidity conditions were less favorable for high $R_0$.

Unlike $R_0$, the effective reproductive number, $R_e$, provides a measure of transmission force during an outbreak; it is calculated as the product of $R_0$ and population susceptibility. Due to higher population susceptibility to the novel strain, the 2009 pandemic had a moderate maximum $R_e$ of 1.24 (1.17–1.29) for the fall wave. Epidemic influenza seasons with the highest $R_e$ were 2003–2004 (1.40, 1.30–1.45), 2007–2008 (1.33, 1.22–1.39), and 2012–2013 (1.30, 1.25–1.35); all of these three seasons experienced moderately severe outbreaks (2, 23, 25).

### Table 1. Parameter estimates by the EAKF and those reported in the literature

<table>
<thead>
<tr>
<th>Season</th>
<th>$S$ maximum, %</th>
<th>$R_0$ maximum</th>
<th>$D$ at maximal $R_0$</th>
<th>$\gamma$ at maximal $R_0$</th>
<th>Attack rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003–2004</td>
<td>70.6 (65.4, 74.1)</td>
<td>2.04 (1.84, 2.21)</td>
<td>1.40 (1.30, 1.45)</td>
<td>4.20 (3.77, 4.50)</td>
<td>1.53 (1.43, 1.63)</td>
</tr>
<tr>
<td>2004–2005</td>
<td>64.6 (60.6, 67.1)</td>
<td>1.98 (1.86, 2.11)</td>
<td>1.25 (1.17, 1.30)</td>
<td>5.00 (4.64, 5.23)</td>
<td>1.66 (1.53, 1.77)</td>
</tr>
<tr>
<td>2005–2006</td>
<td>64.9 (60.7, 68.3)</td>
<td>1.94 (1.86, 2.04)</td>
<td>1.24 (1.15, 1.30)</td>
<td>4.93 (4.35, 5.32)</td>
<td>1.62 (1.54, 1.71)</td>
</tr>
<tr>
<td>2006–2007</td>
<td>64.6 (60.2, 68.2)</td>
<td>1.88 (1.78, 2.01)</td>
<td>1.19 (1.12, 1.24)</td>
<td>5.02 (4.47, 5.39)</td>
<td>1.62 (1.52, 1.73)</td>
</tr>
<tr>
<td>2007–2008</td>
<td>66.8 (62.5, 69.7)</td>
<td>2.03 (1.85, 2.20)</td>
<td>1.33 (1.22, 1.39)</td>
<td>4.97 (4.43, 5.36)</td>
<td>1.57 (1.46, 1.67)</td>
</tr>
<tr>
<td>2008–2009</td>
<td>64.7 (60.7, 68.0)</td>
<td>1.91 (1.78, 2.06)</td>
<td>1.21 (1.16, 1.25)</td>
<td>5.04 (4.68, 5.30)</td>
<td>1.61 (1.52, 1.68)</td>
</tr>
<tr>
<td>2010–2011</td>
<td>64.5 (60.6, 67.2)</td>
<td>1.97 (1.85, 2.11)</td>
<td>1.22 (1.16, 1.27)</td>
<td>4.71 (4.41, 4.93)</td>
<td>1.65 (1.56, 1.76)</td>
</tr>
<tr>
<td>2011–2012</td>
<td>66.3 (62.8, 69.2)</td>
<td>1.86 (1.74, 1.97)</td>
<td>1.18 (1.15, 1.22)</td>
<td>5.18 (4.87, 5.43)</td>
<td>1.58 (1.51, 1.65)</td>
</tr>
<tr>
<td>2012–2013</td>
<td>67.7 (63.5, 70.2)</td>
<td>1.97 (1.84, 2.13)</td>
<td>1.28 (1.21, 1.32)</td>
<td>4.80 (4.13, 5.19)</td>
<td>1.58 (1.43, 1.71)</td>
</tr>
<tr>
<td>2012–2013*</td>
<td>68.9 (65.0, 71.6)</td>
<td>1.96 (1.84, 2.10)</td>
<td>1.30 (1.25, 1.35)</td>
<td>4.76 (4.25, 5.22)</td>
<td>0.62 (0.58, 0.67)</td>
</tr>
</tbody>
</table>

Estimates in this study are segregated by season/pandemic wave over all US cities and shown as mean and interquartile range (in parentheses).

†Corrected estimates for 2012–2013 season, using lower prior values of $\gamma$ (uniform distributions $U(0.33, 2)$ vs. $U(1,10)$ for other seasons) to more accurately account for the elevated flu related search intensity due to media coverage.

‡The entire pandemic (pdm09-10) is defined as the period between weeks ending 11 April 2009 and 2 October 2010.

§The first 2009 pandemic wave (pdm09-w1) is defined as period between weeks ending 11 April 2009 and 22 August 2009.

¶The second 2009 pandemic wave (pdm09-w2) is defined as the period between weeks ending 8 August 2009 and 2 October 2010.
In comparison, the mild 2011–2012 season (26), along with the spring 2009 pandemic wave, had the lowest $R_e$ (Fig. 1D).

The infectious period represents the duration of host infectiousness. It is usually estimated from clinical studies (15) or assumed to follow a certain distribution (e.g., a Gamma distribution) in modeling work (27). Using the ILI+ data and our inference approach, estimates of $D$, the mean infectious period, are substantially lower for the two 2009 pandemic waves (3.79, 3.34±1.7 d) than for the epidemic seasons (4.90, 4.43±2.6 d). This is in agreement with previous studies suggesting a shorter serial interval for the 2009 pandemic strain (15, 20, 28).

Note that estimates of infectious period are larger than past estimates for serial interval as our model does not include a latent period.

Our inference framework includes a parameter, $\gamma$, that maps model-simulated incidence to observed ILI+. This mapping represents the weekly ratio of the probability a person seeks medical attention for any reason to the probability a person seeks medical attention due to influenza (Eq. 4; see Materials and Methods and ref. 5). For the interpandemic seasons during 2003–2012, $\gamma$ was estimated as 1.84 (1.57–1.93) over the flu season. A recent community cohort study (29) suggested a 4.3% probability of seeking medical attention among influenza-infected persons (including those with asymptomatic infection). The CDC estimated 1.25 billion doctor visits per year during 2009–2010 in the United States (30), which converts to a weekly doctor visitation rate of 7.8% of the US population. Combining these two estimates suggests a $\gamma$ of 1.81 (7.8%/4.3%), very close to our model inference estimates.

Because the probability a person seeks medical attention for any reason (i.e., the numerator of $\gamma$) is relatively stationary, the fluctuation in $\gamma$ arises mainly from variation in the probability of seeking medical attention for influenza (i.e., the denominator of $\gamma$). A lower $\gamma$ estimate thus reflects a higher tendency to seek medical help for influenza. The GFT ILI records during the 2012–2013 season are over 2 times CDC estimates, partly due to the intense media coverage accorded influenza that year (23, 31). To account for this issue, we use a broader and lower range of values for the prior of $\gamma$ for the 2012–2013 season than for the other seasons; this choice results in a more reasonable estimate of the attack rate for that season, while estimates for other parameters are similar to those made using the same prior for other seasons (Table 1). Estimated at the time with the maximum epidemic forcing (i.e., maximum $R_e$), the 2012–2013 season had the lowest $\gamma$: the two 2009 pandemic waves and the 2003–2004 season had the next lowest $\gamma$ estimates (Fig. 1F).

These findings are consistent with observations that more people sought medical attention for influenza during the pandemic due to awareness of the ongoing pandemic and during the 2003–2004 and 2012–2013 seasons due to the greater virulence of the strains circulating at the time (24).

Using our estimates of $\gamma$, we are able to convert simulated influenza incidence rates over each flu season into an attack rate. The fall 2009 pandemic wave struck an estimated 27% (22–33%) of the population; the 2003–2004, 2007–2008, 2009–2010, and 2012–2013 seasons had the highest attack rates (19%, 14–24%), and other epidemic seasons had much lower attack rates (10%, 5–13%) (Fig. 1B). These model estimates are based on the final outcomes of infection within the general population, unlike estimates of secondary attack rate that are conditioned on the presence of a primary case within a household (17, 19, 20). The latter method depends on the composition of survey households (e.g., whether a child is present in a household) and thus may not be as representative of the general population.

Our observational measure of influenza incidence, ILI+, is derived in part from municipal-scale GFT ILI estimates. Such municipal-scale ILI+ data are attractive due to their spatial granularity. These GFT estimates, however, have documented problems, including likely statistical overfitting to their CDC ILI target and a tendency to overestimate ILI rates (32). Particularly, the GFT ILI data for the 2012–2013 season were up to 2–3 times CDC reported ILI (23). While combining GFT ILI with regional viral isolation data to create ILI+ helps better reflect influenza activity.
in a local city, systematic observational biases in the ILI+ metric likely still persist.

To examine potential biases introduced by the uncertainty in GFT ILI data, we applied the same model filter frameworks to regional CDC ILI+ data (5) over the same 2003–2013 seasons. Estimates of \( \gamma \) are higher for the simulations using the CDC ILI+ data [e.g., 2.05 (1.69–2.16) vs. 1.84 (1.57–1.93) when using the GFT ILI+ data for the 2003–2012 epidemic seasons]. However, estimates of initial susceptibility, \( R_0 \), maximal \( R_e \), and attack rate are similar using the two data types (see SI Appendix, Table S1).

These findings suggest that the scaling factor, \( \gamma \), in our inference frameworks is compensating for the biases in GFT ILI estimates. The municipal GFT ILI+ data also allow investigation of spatial variations of influenza transmission dynamics among subpopulations within the United States. Of the 115 cities with GFT data, 62 have complete records for all seasons/pandemic waves during 2003–2013. For these 62 cities, we used \( R_0 \) at the time of maximum epidemic forcing (maximum \( R_e \)) as estimated for each of the nine interpandemic outbreaks and two pandemic waves to calculate the correlation of \( R_0 \) between pairs of cities (1,891 pairs in total). This correlation quantifies the chronological covariability of \( R_0 \) during 2003–2013 across cities. In general, cities within the same Health and Human Services (HHS) region tend to have more positively correlated \( R_0 \) estimates (Fig. 2 and SI Appendix, Table S2). Cities along the east coast (i.e., regions 1–4) have highly positively correlated \( R_0 \) (\( r = 0.85 \pm 0.10 \)). Cities within regions 5 and 7 are also highly correlated with cities within regions 1–4 (\( r = 0.83 \pm 0.08 \) among cities within either of these two clusters). In contrast, cities along the west coast (regions 9 and 10) have more distinct \( R_0 \) estimates (\( r = 0.43 \pm 0.35 \) among cities within these two regions; and \( r = 0.32 \pm 0.26 \) among cities within these two regions vs. all others). Four cities—Las Vegas, NV, Phoenix, AZ, Tempe, AZ, and Tucson, AZ—stand out (the blue strip in Fig. 2); \( R_0 \) estimates for these cities tend to vary out of phase with the majority of other cities (\( r = 0.04 \pm 0.27 \)). These cities, particularly Las Vegas and Phoenix, share a similar desert climate, which may contribute to these differences.

**Discussion**

Our findings indicate that online influenza surveillance data can be used in conjunction with a simple dynamical model and data assimilation methods to infer many key epidemiological parameters for both seasonal and pandemic influenza. Moreover, this approach provides joint estimates of these key epidemiological variables and parameters that best match influenza activity over the entire season. This model inference framework is able to depict the evolving epidemiological features of influenza across seasons and over the course of each outbreak.

The inference approach presented here used a simple well-mixed, single-age class SIRS model. The model includes simple distributions for sojourn times in compartments, in particular an exponential distribution for the infectious period, with fixed infectiousness until recovery, two assumptions that are probably not accurate and will bias estimates of the reproductive number (15, 27, 33). Such model misspecification is a potential source of error that suggests caution when interpreting these findings. However, the filters, through their continued adjustment of the model state, are able to partially compensate for...
We did not discriminate strains in the SIRS model. As such, our estimates may confound or blend outcomes due to overlapping outbreaks should there be multiple strains cocirculating. Future studies could address strain-specific inference as strain-specific data become available at the municipal level. Our inference system was run discontinuously for each season; consequently, the immunity period, L, in the SIRS model was less constrained and thus not analyzed here. Constraining an increased number of state variables/parameters introduced by more comprehensive models will require data streams with additional information and finer resolution (e.g., serosurvey on population susceptibility and age-structured surveillance records). Such in-depth inference could be achieved in the future, as data of better quality become available to address these issues. For instance, data with finer age structure may allow more detailed inference on the transmission dynamics of pandemic versus epidemic influenza.

In summary, we have shown that the transmission dynamics of influenza among the general population can be inferred using data assimilation methods and big data estimates of incidence. As more people have access to and increasingly rely on online systems worldwide, mining of similar big data from online social networks may provide valuable information on the early spread of diseases (e.g., the early wave of a pandemic) as well as transmission dynamics for a number of other diseases. Such inference will rely heavily on the quality and reliability of these big data observational estimates.

**Materials and Methods**

**Data.** Weekly ILI+ data are compiled by multiplying weekly municipal GFT ILI, regional GFT ILI, or regional CDC ILI estimates by their corresponding regional influenza viral isolation rate, as reported by the World Health Organization and National Respiratory and Enteric Virus Surveillance System (5). From 2003–2004 to 2012–2013, municipal GFT ILI+ data are available for up to 115 US cities (the number of cities for each year ranges from 66 to 115).

**SIRS Model.** The SIRS model is a well-mixed humidity-forced model that tracks the flow of population in each disease stage by the following equations:

\[
\frac{dN}{dt} = \left( 1 - \frac{S(t)}{N} \right) \frac{R_0}{N} I(t) \\
\frac{dI}{dt} = -\gamma I(t) R(t) \\
\frac{dR}{dt} = \gamma I(t) R(t) - \frac{\xi(t) R(t)}{N} S(t)
\]

where \( N \) is the number of susceptible persons in the population, \( t \) is time, \( N \) is the population size, \( I \) is the number of infectious persons, \( N - I - S \) is the number of resistant individuals, \( \beta(t) \) is the transmission rate at time \( t \), \( L \) is the average duration of immunity, \( D \) is the mean infectious period, and \( \alpha \) is the rate of travel-related import of influenza virus into the model domain. The basic reproductive number at time \( t \) is related to the transmission rate through the expression \( R_0(t) = \beta(t)D, \) and determined by a function with humidity forcing \( \phi(t) \):

\[
R_0(t) = \exp\left( -1800(t + \ln(\text{RH}_{\text{max}} - \text{RH}_{\text{min}})) + \text{RH}_{\text{min}} \right)
\]

where \( \text{RH}_{\text{max}} \) and \( \text{RH}_{\text{min}} \) are, respectively, the maximum and minimum daily basic reproductive number, and \( \phi(t) \) is the specific humidity at time \( t \).

**Mapping the SIRS Model to Observations.** Let \( p(i) \) be the probability of any person contracting influenza during a given week. The ILI+ observation is an estimate of the percentage of influenza visits among all patient visits, or the probability that a person seeking medical attention, \( m \), has influenza, i.e., \( p(i|m) \). By Bayes’ rule, \( p(i) \) is then

\[
p(i|p(m)) = \frac{p(m)p(i|m)}{p(m)} = \frac{p(m)p(i|m)}{p(m)} = p(i|m)\text{ILI}.
\]

On the other hand, the SIRS model simulates the spread of influenza within a perfectly mixed population. For a population of \( N \) people, influenza incidence roughly follows a binomial distribution, i.e., \( B(N, p(i)) \), with an expectancy of \( p(i)N \). That is, in the SIRS model the mean incidence rate, \( \zeta \), is simply \( p(i) \). Accordingly, we derive a model estimate of ILI+, \( \text{ILI}_+, \) using Eq. 4 and the SIRS-simulated incidence rate, \( \zeta \).
where \( y = \frac{\hat{p}(m)}{\hat{p}(m')} \) is a scaling factor mapping simulated influenza incidence occurrence rate, \( \zeta \), to the ILI+ observation (5).

Model—Data Assimilation Methods. We applied either a particle filter with resampling and regularization (PF) (38) or the ensemble adjustment Kalman filter (EAKF) (39) jointly with the SIRS model and ILI+ records. Both model—data assimilation methods are done sequentially by repeated prediction—update cycles. In each cycle, a prediction is made by integrating the SIRS model up to the next observation, and an update is triggered by the arrival of new ILI+ data. The PF simulates the dynamical system with 10,000 randomly generated system replicas, termed particles. The filter then assimilates weekly ILI+ records to evaluate the likelihood of each particle. Particles are selected based on their likelihoods and eventually converge to those with the greatest likelihoods. In comparison, the EAKF generates 300 random system replicas, termed ensemble members. The posterior (i.e., update) of the ensemble mean is weighted based on the prediction (i.e., the prior), the observation, and their respective variances. The EAKF adjusts each ensemble member toward the ensemble mean such that the posterior variance is identical to what is predicted by Bayes’ theorem (4, 5).

Estimation of Key Epidemiological Parameters. We estimate key epidemiology parameters for 2003-2004 through 2012-2013 season for all US cities with ILI+ records using either the PF or the EAKF. Within both filters, each particle/ensemble member includes a set of all state variables and model parameters, which are selected or adjusted at each data assimilation checkpoint; these ensembles provide distributions of each state variable/parameter at each time point. ILI+ data were assimilated from the week ending 11 April 2009 to the week ending 22 August 2009 for the first wave of 2009 pandemic, and from the week ending 8 August 2009 to the week ending 2 October 2010 for the second wave. Simulations for the entire pandemic were made with ILI+ records between weeks ending 11 April 2009 and 2 October 2010. For a given seasonal outbreak, simulations were done from Week 40 of a given year to Week 39 of the next year. We define the week with the maximal and minimal susceptible level, \( S \), as the onset and ending, respectively, of the epidemic/pandemic outbreak, and a flu season as the period between onset and ending. The attack rate is calculated as the sum of simulated incidence from onset to ending. The timing of maximum epidemic forcing is defined as the week with the highest effective reproductive number. The basic reproductive number \( R_0 \) effective reproductive number \( R_e \), and infectious period \( D \) are estimated at the week with the maximum epidemic forcing.

ACKNOWLEDGMENTS. Funding was provided by US National Institutes of Health (NIH) Grants GM100467 (to W.Y., M.L., and J.S.) and GM110748 (to J.S.) and the NIH Models of Infectious Disease Agent Study program through Cooperative Agreement U54GM088558 (to J.S. and M.L.), as well as National Institute of Environmental Health Sciences (NIHES) Center Grant ES00908 (to J.S.) and the Research and Policy for Infectious Disease Dynamics (RPIDD) Program of the Science and Technology Directorate, US Department of Homeland Security (J.S.).