Epidemics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Geographic transmission hubs of the 2009 influenza pandemic in the United States

Stephen M. Kissler[a,*], Julia R. Gog[a], Cécile Viboud[c], Vivek Charu[c,d], Ottar N. Bjørnstad[e], Lone Simonsen[c,f], Bryan T. Grenfell[b,c]

[a] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge, United Kingdom
[b] Department of Ecology and Evolutionary Biology, University of Princeton, Princeton, NJ, USA
[c] Fogarty International Center, National Institutes of Health, Bethesda, MD, USA
[d] Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[e] Department of Entomology, Pennsylvania State University, University Park, PA, USA
[f] Department of Public Health, University of Copenhagen, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Pandemic influenza
Transmission hubs
Metapopulation
Gravity model
Phylogeography

ABSTRACT

A key issue in infectious disease epidemiology is to identify and predict geographic sites of epidemic establishment that contribute to onward spread, especially in the context of invasion waves of emerging pathogens. Conventional wisdom suggests that these sites are likely to be in densely-populated, well-connected areas. For pandemic influenza, however, epidemiological data have not been available at a fine enough geographic resolution to test this assumption. Here, we make use of fine-scale influenza-like illness incidence data derived from electronic medical claims records gathered from 834 3-digit ZIP (postal) codes across the US to identify the key geographic establishment sites, or "hubs", of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the United States. A mechanistic spatial transmission model is fit to epidemic onset times inferred from the data. Hubs are identified by tracing the most probable transmission routes back to a likely first establishment site. Four hubs are identified: two in the southeastern US, one in the central valley of California, and one in the midwestern US. According to the model, 75% of the 834 observed ZIP-level outbreaks in the US were seeded by these four hubs or their epidemiological descendants. Counter-intuitively, the pandemic hubs do not coincide with large and well-connected cities, indicating that factors beyond population density and travel volume are necessary to explain the establishment sites of the major autumn wave of the pandemic. Geographic regions are identified where infection can be statistically traced back to a hub, providing a testable prediction of the outbreak's phylogeography. Our method therefore provides an important way forward to reconcile spatial diffusion patterns inferred from epidemiological surveillance data and pathogen sequence data.

1. Introduction

Recent years have seen a surge in the use of mathematical models to describe the geographic transmission of infectious diseases (Colizza et al., 2007b; Tatem, 2014; Riley et al., 2015). A central goal is to identify "hotspots" of spread, be they individuals who disproportionately contribute to transmission within a population (Galvani and May, 2005; Paull et al., 2012), or cities or countries that act as sites where an epidemic first becomes established (that is, where sustained chains of transmission first take hold) (Legrand et al., 2009; Levy et al., 2011; Yang et al., 2015b). Identifying hotspots can help guide surveillance and intervention efforts that may prevent or slow the spread of disease transmission (Skene et al., 2014; Russell et al., 2008). An important type of geographic hotspot is the 'hub', which is a site of epidemic establishment that contributes substantially to the onward geographic spread of a disease. Specifically, a hub may be defined as a location where (a) an outbreak occurs due to the establishment of a long-distance pathogen introduction from outside the population, which then (b) contributes significantly to onward spatial spread within the population. Hubs may be contrasted with 'sources', which are sites where a new genetic variety of a pathogen first emerges (Viboud et al., 2013), and also from 'superspreaders', which are sites that spread infection to many immediate neighbors – though all three may sometimes coincide. Transmission hubs are often thought to coincide with locations with high connectivity and population density (Xia et al., 2004; Ferguson et al., 2006). For influenza, however, this association has not

* Corresponding author.
E-mail address: sk792@cam.ac.uk (S.M. Kissler).

https://doi.org/10.1016/j.epidem.2018.10.002
Received 3 August 2017; Received in revised form 5 October 2018; Accepted 8 October 2018
1755-4365/ © 2018 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

been tested, nor has any data-validated mechanistic identification of transmission hubs at the country scale been performed.

The increased availability of geo-tagged epidemiological data now makes such a mechanistic analysis possible. In particular, electronic medical claims records (EMRs) offer a so-far underutilized source of large-volume syndromic influenza-like illness (ILI) data. Though the specificity of EMRs for diagnosing influenza is lower than the specificity of laboratory-based tests, EMRs give reliable estimates of influenza epidemic timing, especially when compared with data derived from social media and Internet search platforms like Twitter and Google (Viboud et al., 2014; Lampos et al., 2010; Carneiro and Mylonakis, 2009; Olson et al., 2013). EMRs are available at a high enough geographic resolution to perform robust statistical analyses on influenza transmission at a range of spatial scales (Viboud et al., 2014; Gog et al., 2014).

To trace chains of infection geographically, it is necessary to infer the relative onset times of local outbreaks. For time series data, a range of outbreak detection algorithms exist. However, most have been developed to either detect the presence of an epidemic during some time interval (but not necessarily its precise timing) (Held et al., 2006; Pelecanos et al., 2010), or to provide early warning of an unfolding outbreak (Hashimoto et al., 2000; Wagner et al., 2001; Abeku et al., 2004; Won et al., 2017). Few exist that are specifically tailored to detect epidemic onset time after an outbreak has occurred (see Gog et al., 2014; Charu et al., 2017 for two examples), possibly because epidemiological data at a sufficiently detailed geographic resolution have so far not been available to motivate this sort of inference. There remains a need to refine and systematically compare the outbreak onset detection algorithms that do exist, to determine which most reliably detects outbreak onset times from retrospective time series data.

Influenza genetic sequences offer an alternative data stream for inferring geographic disease transmission patterns. Integrating epidemiological and phylogenetic inferences is an important emerging area of research (Grad and Lipsitch, 2014). Some previous studies have successfully combined these data streams at the global scale (Kilpatrick et al., 2006; Lycett et al., 2012), but at the continental scale and smaller, the task has proven difficult, possibly due to a lack of genomic data at sufficient geographic resolution (Viboud et al., 2013).

In this article, we identify key drivers of geographic transmission of the autumn wave of the 2009 A/H1N1pdm influenza outbreak in the United States using a mechanistic mathematical model, based on Gog et al. (2014). The model is fit to outbreak onset times inferred from medical insurance claims data collected during outpatient visits in 834 3-digit ZIP codes across the US. We then apply a Markov strategy to probabilistically trace the inferred transmission routes backwards in time and identify the transmission hubs of the outbreak. Tracing onward spread from the hubs yields a testable prediction for the phylogeographic structure of the outbreak, opening the possibility of refining the inference of geographic transmission chains and transmission hubs using combined genetic and epidemiological data.

2. Methods

2.1. Data

Data for this analysis come from a convenience sample of electronic medical claims forms (type CMS-1500) submitted by primary care physicians across the US and maintained by IMS Health (originally SDI Health). Each claim is associated with a single outpatient visit, and includes one or more ICD-9 codes (Moriyama et al., 2011) listed by the physician that describe the patient's illness. The overall sample is thought to capture over 50% of all outpatient visits in the US in 2009 (Viboud et al., 2014). The records are binned weekly, and aggregated geographically by the first three digits of the ZIP (postal) code of the practice from which they are submitted (U.S. Postal Service Office of Inspector General, 2013). These three-digit ZIP codes will be referred to simply as 'ZIPs' (not to be confused with the finer five- or ten-digit ZIP codes, also assigned to many mailing addresses in the US (Moriyama et al., 2011)). Time series of weekly influenza-like illness (ILI) incidence are created by extracting claims with a direct mention of influenza, or fever combined with a respiratory symptom, or febrile viral illness (ICD-9 487-488 OR [780.6 and (462 or 786.2)] OR 079.99), following Viboud et al. (2014). For each ZIP, the number of ILI cases in each week is divided by the total number of patients who visited a physician in that ZIP during that week, yielding an 'ILI ratio' time series. There are 884 ILI ratio time series, one for each ZIP in the lower 48 US states, each spanning 52 weeks from the week commencing 4 Jan 2009 through the week commencing 27 Dec 2009. 50 ZIPs are excluded from further analysis due to excessive noise (see 'Definition of epidemic onset'). The remaining 834 time series will be referred to as the IMS-ILI dataset, following Viboud et al. (2014).

The correspondence between the IMS-ILI dataset and reference influenza surveillance data from the US Centers for Disease Control and Prevention (CDC) is described in depth by Viboud et al. (2014). In brief, the weekly incidence and peak timing of outbreaks in the IMS-ILI dataset both correlate highly with the weekly incidence and peak timing from CDC ILI and CDC virologic surveillance data at the regional level. Correlations for both metrics remain strong when the IMS-ILI data are compared with city-level ILI data from New York State. City-level correlations outside New York State could not be assessed, due to a lack of reference data. Taken together, this suggests that the IMS-ILI data provide reliable information about epidemic timing by geographic region in the US (Viboud et al., 2014).

Population sizes and geographic coordinates for the ZIPs are available from the US Census Gazetteer files (United States Census Bureau, 2015). The Gazetteer files partition data into finer five-digit postal codes, so the three-digit ZIP population sizes are calculated by summing the population sizes of all constituent five-digit codes, and the three-digit ZIP coordinates are obtained by taking the population-weighted mean of the coordinates of the constituent five-digit codes. There are 21 ZIPs with population size smaller than 20,000 that we omit from the analysis (see Fig. S2). City names for each ZIP are available from the United States Postal Service (United States Postal Service, 2018). These names are not necessarily unique: a single name may apply to multiple ZIPs, since a city can be partitioned into multiple ZIPs. We will always refer to ZIPs both by their three-digit number, which is a unique identifier, and by their associated city name.

Data on school start dates are available at the state level from Chao et al. (2010). In Alabama, Florida, Georgia, Mississippi, and South Carolina, the five states near the apparent epicenter of the outbreak in the eastern US, school start dates are available at the finer district level (also from Chao et al. (2010)). Most ZIPs contain multiple school districts, so we define the ZIP-level school start date to be the median of all district start dates within that ZIP. We repeated the full analysis below using the earliest, rather than the median, district-level school start date within each ZIP; the form of the optimal transmission model and the final set of transmission hubs were unchanged.

2.2. Definition of epidemic onset

At least two strategies exist for explicitly inferring outbreak onset times from retrospective ILI incidence time series (Charu et al., 2017; Gog et al., 2014). The first strategy, the "threshold method" from Gog et al. (2014), defines the onset time of an influenza outbreak as the first of three consecutive weeks in which ILI incidence in a given location surpasses a sinusoidal baseline fit to the ILI incidence between influenza seasons. This is similar in strategy to many outbreak detection methods used for real-time epidemic surveillance (Abeku et al., 2004; Hashimoto et al., 2000; Wagner et al., 2001; WHO Global Influenza Programme Surveillance and Epidemiology team, 2012; Won et al., 2017) and is closely related to the threshold method used by Eggo et al. (2011). Though conceptually straightforward, defining baselines and thresholds

can be difficult and somewhat *ad hoc* (Centers for Disease Control and Prevention, 2016; Shmueli and Burkom, 2010). The second strategy, the "breakpoint method" introduced by Charu et al. (2017), takes a fundamentally different approach. It estimates epidemic onset time as the changepoint in the slope of a bi-linear trend fit to an ILI time series in the n weeks preceding the epidemic peak, where n is a parameter chosen by the modeler. This avoids the need to define a baseline, and provides a natural way of characterizing uncertainty in the onset estimate (see SI §1.1). A full mathematical specification of the breakpoint method is given in SI §1. A simulation-based side-by-side evaluation of the breakpoint method and a version of the threshold method adapted from Gog et al. (2014) demonstrates that the breakpoint method yields onset estimates with greater accuracy and precision than the threshold method (see SI §1.2).

To infer ZIP-level outbreak onset times from the 2009 IMS-ILI data, the breakpoint method is implemented by fitting a bi-linear trend to each ZIP's ILI ratio time series in the $n = 17$ weeks prior to and including the week of peak incidence in that ZIP. The week of peak incidence is defined as the week in which the maximum ILI ratio is reached, between 5 July 2009 and 27 Dec 2009. The choice of 17 weeks provides enough data points to give a robust onset estimate, while avoiding the tail end of the spring infection wave that affected a number of locations. The onset date is defined as the maximum likelihood estimate of the breakpoint in the bi-linear trend, rounded to the nearest half-week, following Gog et al. (2014) and Charu et al. (2017). Uncertainty is assessed using the log-likelihood profile of the breakpoint onset estimate. ZIPs with uncertain onset times, defined here as those for which the log-likelihood profile of the breakpoint estimate does not drop by at least 2 units, undergo a second fitting procedure to determine whether a more precise onset estimate might be obtained using 'alternate' peaks (see SI §1.1). If the breakpoint log-likelihood profile still does not drop by at least 2 units after this adjustment, the ZIP is omitted from further analysis. There are 29 of these omitted ZIPs, depicted geographically in Fig. S2. This leaves 834 ZIPs for further analysis. The breakpoint method is illustrated in Fig. 1 on the ILI ratio time series for ZIP 606 (Chicago IL) and the nearby ZIP 538 (Madison WI). Breakpoint onset times for the 834 ZIPs are depicted geographically in Fig. 2. An exploratory analysis of the breakpoint onset times and their uncertainties is presented in SI §1.3.

2.3. Mechanistic transmission model

The transmission model considered here is based on the most parsimonious model selected by Gog et al. (2014). It is defined as

$$\lambda_i(t) = \beta_0 + (\beta_d + I\beta_{ds})N_i^{\mu}\frac{\sum_{j\in\Lambda} N_j^{\nu}\kappa(d_{ij})}{\left[\sum_{j\neq i} N_j^{\nu}\kappa(d_{ij})\right]^{\epsilon}} \quad (1)$$

where $\lambda_i(t)$ is the force of infection on location i at time t; I is an indicator function that is 1 if schools are in session in location i at time t and is 0 otherwise; N_i and N_j are the population sizes of locations i and j, divided by average population size over all locations; Λ is the set of infected locations at time t; and d_{ij} is the great circle distance in kilometers between locations i and j. Time t is treated as a discrete variable, with units of one half week. The parameter β_0 accounts for force from external seeding; β_d and β_{ds} together define a local transmission factor that is modulated by schools being in session; μ and ν define how the force of infection relates to the population sizes of the recipient and donor locations, respectively; and ϵ adjusts the normalization term accounting for population density around location i. The function $\kappa(d_{ij})$ is a kernel describing how the force of infection decays with the distance between ZIPs. Two kernels are tested: a power kernel, with form

$$\kappa(d_{ij}) = d_{ij}^{-\gamma}, \quad (2)$$

and an exponential kernel, with form

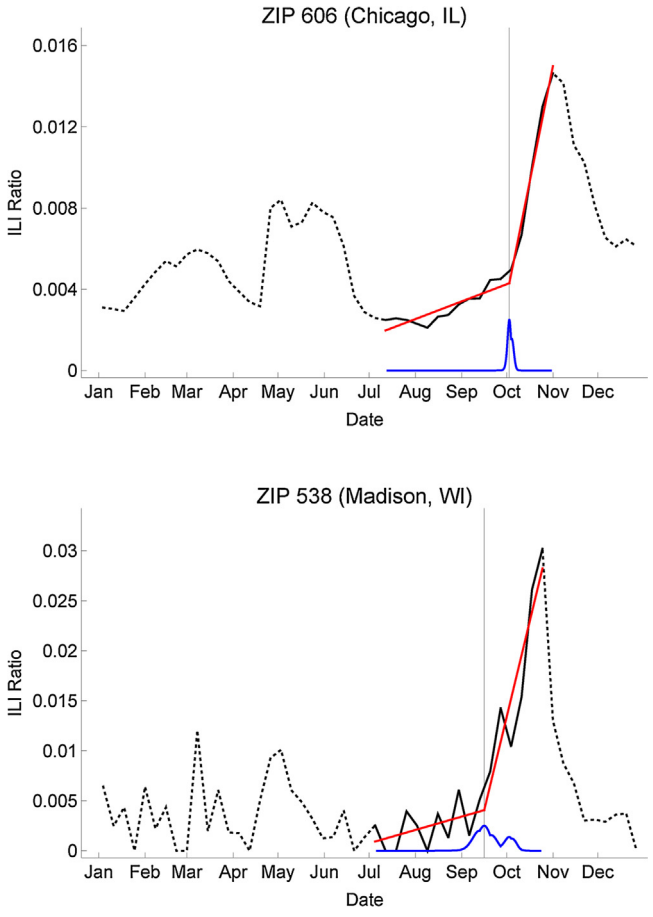

Fig. 1. Illustration of the breakpoint method for determining time of epidemic onset. These time series (dotted lines) depict the weekly ILI ratios for ZIP 606 (Chicago IL) and the nearby ZIP 538 (Madison WI) in 2009. ZIP 606 is one of the largest ZIPs considered (pop. 2.8 million), and ZIP 538 is one of the smallest (pop. 58,000). A bi-linear trend (red) is fit to the 17 weeks of the time series (solid black) prior to and including the week of peak incidence. The onset date is defined as the maximum likelihood estimate of the breakpoint in the bi-linear trend, rounded to the nearest half-week. The blue curve below the time series depicts the likelihood profile for the breakpoint onset. This curve describes uncertainty in the breakpoint estimate. For ZIP 606, the distribution is narrow, indicating a high degree of certainty in the onset estimate. For ZIP 538, the distribution is wider and bimodal. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

$$\kappa(d_{ij}) = \mathrm{Exp}[-d_{ij}/\rho], \quad (3)$$

where γ and ρ are parameters that describe how quickly the kernel decays with distance. Power kernels generally decay less quickly at long distances than exponential kernels (see Fig. S14), so we expect that a power kernel will capture transmission dynamics with many long-distance jumps of infection, while an exponential kernel will capture highly localized transmission dynamics. The transmission model given by Eq. (1) is a version of the 'gravity model', since it describes the coupling between geographic locations as a function of their population sizes and the distance between them (Gog et al., 2014; Wilson, 1970; Eggo et al., 2011; Truscott and Ferguson, 2012).

Given model parameters $\Theta = \{\beta_0, \beta_d, \beta_{ds}, \mu, \nu, \gamma, \epsilon\}$ or $\Theta = \{\beta_0, \beta_d, \beta_{ds}, \mu, \nu, \rho, \epsilon\}$, the probability that a location i becomes infected at time T_i is

$$P_i(T_i|\Theta) = (1 - e^{-\lambda_i(T_i)})\prod_{t=1}^{T_i-1} e^{-\lambda_i(t)} \quad (4)$$

following Eggo et al. (2011). Hence, the probability of observing the full set of onsets $T = \{T_1, T_2, \ldots, T_n\}$ is a product across all n locations:

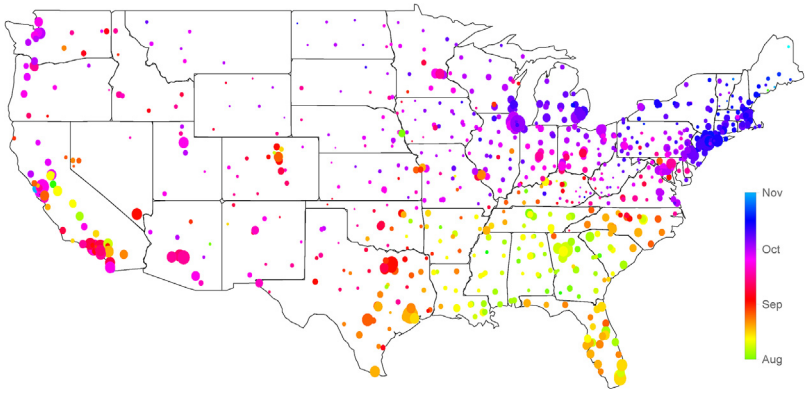

Fig. 2. Epidemic onsets for the 834 three-digit ZIP codes. Disc area is proportional to the ZIP's population size. The earliest outbreaks are depicted in green/yellow, and the latest in purple/blue. A major epidemic wave emanated out of the southeastern United States, with a possible second seeding event in California. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

$$P(T|\Theta) = \prod_{i=1}^{n}\left[(1 - e^{-\lambda_i(T_i)})\prod_{t=1}^{T_i-1} e^{-\lambda_i(t)}\right]. \quad (5)$$

2.3.1. Model fit

Model parameters are estimated using maximum likelihood. The log-likelihood of the model parameters Θ given the breakpoint epidemic onset times $T = \{T_1, \ldots, T_n\}$ for locations 1, …, n is

$$\ell(\Theta; T) = \sum_{i=1}^{n}\left(\log(1 - e^{-\lambda_i(T_i)}) - \sum_{t=1}^{T_i-1}\lambda_i(t)\right). \quad (6)$$

The parameter values that maximize Eq. (6) are calculated using the Nelder-Mead simplex algorithm, as implemented in MATLAB's `fminsearch()`. The fits of nested sub-models, obtained by setting various parameters to 'null' values and re-fitting the parameters, are compared using the Akaike Information Criterion (AIC) (Burnham and Anderson, 2004), which finds an optimal tradeoff between fit and parsimony by rewarding models with higher likelihood and penalizing models with many parameters. Models with lower AIC are considered more optimal. Table 1 gives the relative AIC values for a representative set of nested models derived from Eq. (1). For the models represented in Table 1, β_{ds} and ν are fixed at their null values of 0, since this always yields a more optimal model (see Table S3). Table S3 gives values of the AIC values for all possible nested models. The optimal model in terms of AIC uses the exponential kernel with full density dependence ($\epsilon = 1$), and does not select effects from schools (β_{ds}) or from donor population size (ν). The resulting model is

$$\lambda_i(t) = \beta_0 + \beta_d N_i^{\mu}\frac{\sum_{j\in\Lambda} e^{-d_{ij}/\rho}}{\left[\sum_{j\neq i} e^{-d_{ij}/\rho}\right]}. \quad (7)$$

Parameter values for this model are given in Table 2. A validation of this model, including forward epidemic simulations and an evaluation of the model's sensitivity to onset detection method, spatial resolution, and perturbations in epidemic onset times, is provided in the SI §3.

Table 1
ΔAIC values for nested sub-models of the geographic transmission model, Eq. (1). Sub-models represented here allow ϵ and μ to be either free or fixed at a null value, and fix $\beta_{ds} = \nu = 0$, using the power (Eq. (2)) and exponential (Eq. (3)) kernels. The optimal model is normalized to have a ΔAIC of 0.

ϵ	μ	ΔAIC Power	ΔAIC Exponential
Free	Free	281.0	0.5
Free	0	320.4	36.0
1	Free	311.7	**0**
1	0	380.1	38.6
0	Free	599.6	345.0
0	0	606.9	366.5

Table 2
Fitted parameter values for the optimal transmission model, Eq. (7).

Parameter	Description	Units	Value	95% confidence interval
β_0	Background transmission rate	$(\Delta t)^{-1}$	0.00040	(0.00014, 0.00083)
β_d	Spatial transmission coefficient	$(\Delta t)^{-1}$	0.77	(0.71, 0.84)
μ	Exponent of dependence on recipient population size	None	0.23	(0.16, 0.30)
ρ	Characteristic distance for the exponential kernel	km	96	(85, 110)

2.4. Identification of epidemic hubs

A key property of the transmission model is that, for a given ZIP i, the force of infection contributed by each ZIP j is additive, and therefore separable. Define $\lambda_{i,j}$, the force on ZIP i from ZIP j at i's time of onset (T_i), as

$$\lambda_{i,j} = \begin{cases} \beta_d N_i^{\mu}\frac{e^{-d_{ij}/\rho}}{\left[\sum_{m\neq i} e^{-d_{im}/\rho}\right]} & \text{if } T_j < T_i \\ 0 & \text{otherwise} \end{cases}$$

The total force of infection on ZIP i at its time of onset can be written as the sum of these components:

$$\lambda_i(T_i) = \beta_0 + \sum_{j=1}^{n}\lambda_{i,j}.$$

This is equivalent to the transmission model given in Eq. (7).

The independent contributions $\lambda_{i,j}$ can be visualized as a transmission network, as depicted in Fig. 3A. ZIPs are represented by nodes (circles), connected with arrows that indicate possible transmission pathways. In addition, n 'seeding states' (clouds) are introduced, each of which exerts a force of β_0 on a single ZIP. Conceptually, there is really just one common source of external seeding that exerts a constant force of β_0 on all ZIPs. However, as discussed later in this section, this strategy of separating the force from external seeding into a unique state for each ZIP makes it possible to infer the probability that a specific external seeding event caused any particular outbreak.

To identify epidemic hubs, transmission chains are traced back to their point of first introduction. This is done by reversing the direction of the transmission network and noting that, with the proper normalizations, the resulting 'reverse transmission network' represents a Markov chain for which the probability of transitioning from state i to state j is equivalent to the probability that ZIP i was infected by 'parent' ZIP j. From this perspective, stepping backwards through the transmission network is equivalent to taking subsequent powers of the Markov chain's transition matrix. In the limit as these powers approach infinity, the i, jth entry of the exponentiated transition matrix gives the

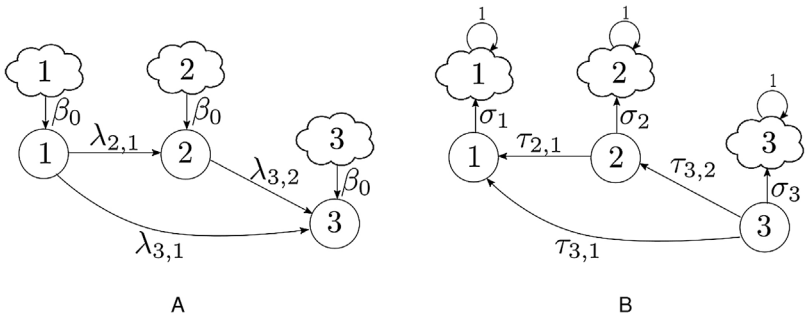

Fig. 3. Forward transmission network (A) and reverse transmission network (B) for an idealized outbreak taking place on three ZIPs. Circles represent ZIPs, and clouds represent 'seeding states' that capture infective force that cannot be explained by gravity-driven local transmission. The outbreak begins in ZIP 1, then infects ZIP 2, and finally infects ZIP 3, in three subsequent time steps. In Diagram A, arrows denote possible transmission paths, and arrow labels give the infective force at time of infection. In Diagram B, arrows point toward possible 'parent' outbreaks, and arrow labels give the probability that the ZIP at the tip of the arrow directly sparked the outbreak in the ZIP at the tail of the arrow. Definitions of the arrow weights are given in Section 2. In this simplified setting, ZIP 1 would be a hub, since the outbreaks in ZIPs 2 and 3 can be traced back to the seeding state attached to ZIP 1 with high probability. The hub calculation procedure accounts for ZIP 1's direct influence on ZIPs 2 and 3, as well as its indirect influence on ZIP 3 via ZIP 2.

probability that the outbreak in ZIP i was initially triggered by a seeding event in ZIP j, under the assumed transmission model.

To illustrate the procedure, refer again to the idealized outbreak depicted in Fig. 3. Reversing the arrows in Diagram A gives the reverse transmission network (Diagram B), where each arrow now points toward a possible contributor of infection. The transition probabilities are denoted

$$\tau_{i,j} = P(\text{transmission from } j \text{ to } i | i \text{ is infected at time } T_i) = \frac{\lambda_{i,j}}{\lambda_i(T_i)}$$

and

$$\sigma_i = P(\text{external seeding in } i | i \text{ is infected at time } T_i) = \frac{\beta_0}{\lambda_i(T_i)}.$$

The $\tau_{i,j}$ represent the probability that the outbreak in ZIP i came from parent ZIP j, and the σ_i represent the probability that the outbreak in ZIP i was directly due to a seeding event.

Define $\boldsymbol{\tau}_{n\times n}$ to be the matrix whose i, jth entry is $\tau_{i,j}$. Note that $\tau_{i,j} = 0$ for all $j \geq i$, so $\boldsymbol{\tau}$ is strictly lower triangular. Also define $\boldsymbol{\sigma}_{n\times n}$ to be the matrix with σ_1, σ_2, …, σ_n along the diagonal and with zeros elsewhere. The transition matrix $\mathbf{M}_{2n\times 2n}$ that describes the reverse transmission network can be written using these matrices:

$$\mathbf{M} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \boldsymbol{\sigma} & \boldsymbol{\tau} \end{pmatrix}.$$

The first n elements of the state space of $\mathbf{M}$ correspond to the seeding states (clouds in Fig. 3), and the remaining n elements correspond to the ZIPs. Entry $\mathbf{M}_{i,j}$ is the probability that 'parent' element j directly sparked element i's outbreak (or, equivalently, the probability that the reverse transmission process transitions from element i to element j). The identity matrix in the upper left block indicates that the seeding states are ultimate sources of infection; they can only transition to themselves. Similarly, the $\mathbf{0}$ matrix in the upper right block indicates that transmission cannot occur from a ZIP to a seeding state. The $\boldsymbol{\sigma}$ matrix in the lower left block captures the probability of a seeding event in each ZIP. The $\boldsymbol{\tau}$ matrix in the lower right captures the transmission probabilities between ZIPs. Note that, as required, the row sums of $\mathbf{M}$ all equal 1.

The pth power of $\mathbf{M}$ contains the probabilities of transitioning between any two nodes via $p-1$ intermediate steps. Finding the ultimate ancestor of each location's outbreak, then, requires calculating $\lim_{p\to\infty} \mathbf{M}^p \equiv \mathbf{M}^\infty$. Since $\boldsymbol{\tau}$ is strictly lower triangular (has zeros along its diagonal), $\boldsymbol{\tau}^m = \mathbf{0}$ for $m \geq n+1$. Thus, $\mathbf{M}^m = \mathbf{M}^\infty$ for $m \geq n+1$, yielding

$$\mathbf{M}^\infty = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ (\mathbf{I} + \boldsymbol{\tau} + \boldsymbol{\tau}^2 + \ldots)\boldsymbol{\sigma} & \mathbf{0} \end{pmatrix}.$$

Element $(\mathbf{M}^\infty)_{i,j}$ gives the probability that seeding state j was the ultimate source of the outbreak in ZIP i. The identity matrix in the upper left block indicates that seeding states are sources unto themselves, as constructed. Each ZIP's ultimate source is a seeding state, since the ZIP → ZIP transitions in the lower-right block of the matrix all go to zero. The lower-left block of $\mathbf{M}^\infty$ contains the values of greatest interest. Denote this block $\mathbf{P}_{n\times n} \equiv (\mathbf{I} + \boldsymbol{\tau} + \boldsymbol{\tau}^2 + \ldots)\boldsymbol{\sigma}$. The entries $\mathbf{P}_{i,j}$ are the probabilities that external seeding (force from β_0) in ZIP j ultimately led to an outbreak in ZIP i. The row sums $\Sigma_j \mathbf{P}_{i,j}$ equal 1 for all j. The column sums of $\mathbf{P}$, denoted $C_j = \Sigma_i \mathbf{P}_{i,j}$, can be interpreted as the expected number of outbreaks triggered by seeding in ZIP j. Any ZIP i for which the associated seeding state has $C_i > 41.7 = 0.05N$, where $N = 834$ is the number of ZIPs (i.e. seeding in the location triggered outbreaks in effectively 5% or more of the observed locations), *and* where $\sigma_i > 0.3$ (i.e. there is a greater than 30% chance that the location's outbreak was caused by external seeding), is classified as a hub.

3. Results

3.1. Drivers of local transmission

The optimal transmission model (Eq. (7)) includes significant effects from recipient population size, population density, and geographic distance. Maximum likelihood parameter values for this model are listed in Table 2. In agreement with Gog et al. (2014), donor population size is not selected as a significant driver of geographic influenza transmission (that is, ν is fixed at zero in the optimal model). In contrast to Gog et al. (2014), school start dates are not selected as key drivers of geographic influenza transmission (that is, β_{ds} is fixed at zero in the optimal model), though this finding should be interpreted with care (see Section 4).

3.2. Epidemiological coupling between cities decays exponentially with distance

A power law is the most common choice of distance kernel for the gravity model (Zipf, 1946; Wilson, 1970; Truscott and Ferguson, 2012). However, for this analysis of the autumn 2009 A/H1N1pdm influenza outbreak in the United States, a more quickly-decaying exponential kernel provides a significantly better fit (see Table 1 and Table S3). This reinforces the importance of short-range transmission over long-distance jumps of infection in the geographic spread of the autumn 2009 A/H1N1pdm influenza pandemic wave.

3.3. Hubs of the autumn wave of the 2009 A/H1N1pdm influenza pandemic in the US

We identify four transmission hubs of the autumn 2009 A/H1N1pdm influenza outbreak in the United States. The two hubs that triggered the greatest effective number of onward infections lie in the southeastern US, in Georgia and Mississippi. The hubs that triggered the third and fourth greatest effective number of onward infections lie in the central valley of California and in eastern Nebraska, respectively. The four hubs, the effective number of outbreaks C that each triggered, and associated information are listed in Table 3, and are depicted geographically in Fig. 4 (boxes).

Table 3
Hubs of the autumn 2009 A/H1N1pdm pandemic wave in the United States. Columns give the ZIP number, name, population size, effective number of onward outbreaks triggered (C), probability that the hub's outbreak was due to external seeding (σ), and the outbreak onset date as estimated by the breakpoint method, for each hub.

ZIP	Name	Pop. size	C	σ	Onset date
389	Grenada, MS	113,782	339.3	1.00	23 Jul
398	Albany, GA	111,263	155.0	0.55	26 Jul
952	Stockton, CA	508,759	78.9	1.00	26 Jul
681	Omaha, NE	573,828	51.3	0.99	2 Aug

The forward transmission triggered by each hub can be mapped geographically. Fig. 4 depicts the probability that each ZIP's outbreak came from any given hub. Each hub is assigned a color, and then each ZIP is assigned a hue in proportion to the probability that its outbreak came from that hub. The colors are allowed to mix, to capture mixed influence from multiple hubs. Fig. 4 therefore illustrates the most likely forward coalescence of the transmission chains from each hub. For example, the ZIP-level outbreaks in California can all be traced back with high probability to the hub in Stockton, CA. In the northeast, on the other hand, the most likely points of origin are Grenada, MS and Albany, GA, as indicated by the purple hue. The results underlying this map offer a link between population-level and genetic-level influenza data streams. It is reasonable to believe that outbreaks in locations with similar hues may have been colonized by genetically-related viruses, due to the likelihood of the virus' common point of origin. The inferred transmission chains may therefore reflect the phylogeographic structure of the autumn 2009 H1N1pdm outbreak in the United States. This structure could be tested and refined with sufficiently resolved genetic data.

Though four ZIPs are pinpointed here as transmission hubs of the autumn 2009 A/H1N1pdm influenza outbreak, the precise identity of the hubs is subject to some uncertainty. Re-identifying the hubs using onset times randomly drawn from the breakpoint likelihood profiles, rather than the maximum likelihood onset times, sometimes yields a different set of hubs. Normally, these new hubs lie geographically close to the four hubs identified here (see Fig. S39). When accounting for onset uncertainty, ZIPs 389 (Grenada), 398 (Albany), and 952 (Stockton) are still almost always chosen as hubs, and when they are not, the hubs that replace them still normally lie in the southeastern US and in the central valley of California (see SI §6). So, while we cannot say definitively that the 2009 A/H1N1pdm influenza pandemic first became established in the four ZIPs listed in Table 3, there is good evidence that there were important establishment events in the southeastern US and in the central valley of California.

It does not appear that the hubs of the 2009 A/H1N1pdm influenza pandemic were in major cities, contrasting with the hypothesis that epidemic establishment should normally take place in well-connected and densely-populated areas. This counter-intuitive distribution of hubs does not appear to be due to a bias in data or methods. One might expect ZIPs with smaller population sizes to have noisier ILI time series, which might cause the onset detection method to mistakenly identify earlier epidemic onsets in smaller ZIPs than in larger ones. A scatter plot depicting epidemic onset vs. population size, however, reveals little relationship between onset time and ZIP population size (Fig. S7). The linear regression trend line has a negative slope ($R = -0.21$, $p < 10^{-4}$), suggesting that smaller ZIPs tend to have later, not earlier, onsets than larger ZIPs. Also, plotting the 95% onset confidence interval vs. population size for all ZIPs shows that there is no relationship between ZIP population size and onset uncertainty (Fig. S9). The largest onset uncertainties are observed in Los Angeles, San Francisco, and New York City (Fig. S8). Finally, for simulated outbreaks, the breakpoint method reliably estimates epidemic onset time even when the epidemic time series is noisy (see Section 2 and Fig. S6). So, it is unlikely that the observed set of hubs can be explained by artifacts from the data.

Furthermore, the transmission hubs are fairly robust to geographic data resolution. The IMS-ILI data are also available at a coarser spatial resolution, aggregated geographically by Sectional Center Facility (SCF), rather than by 3-digit ZIP. The SCF-level data have been considered previously (Gog et al., 2014; Charu et al., 2017). At the SCF level, Grenada MS, Albany GA, and Stockton CA are again identified as transmission hubs with the same relative ordering in C, while Omaha NE is replaced by Ashland KY (see SI §6).

4. Discussion

This is the first study to infer the sites of establishment from long-distance introductions of pandemic A/H1N1pdm influenza virus within the United States using a mechanistic epidemiological model. This was made possible by fine-scale medical claims data made available by IMS Health, and by an improved geographic transmission model for the spread of the autumn 2009 A/H1N1pdm influenza outbreak in the US. These refinements respectively provided the precision and accuracy needed to identify the hubs of the autumn wave of the 2009 pandemic in the US. Our method follows in the spirit of previous work that aims to locate the transmission hubs of an epidemic (Legrand et al., 2009; Levy et al., 2011; Yang et al., 2015b), but has the advantages that it accounts for the actual sequence of infected cities rather than assuming general diffusive spread, has guaranteed convergence once the underlying transmission model is specified, and does not assume a fixed number of introduction sites *a priori*.

Contrasting with conventional wisdom, all four hubs lie in mid-sized cities, not in the highly-connected urban centers that are often associated with outbreak establishment. While it is likely that air travel played an important role in disseminating the 2009 A/H1N1pdm virus both internationally and within the US during the early spring wave

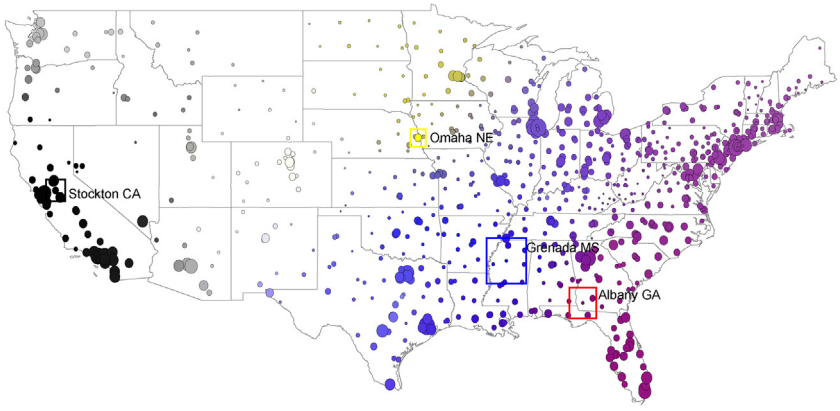

Fig. 4. Basins of infection for the four hubs listed in Table 3. Hubs are outlined with boxes. Box area is proportional to the number of outbreaks that seeding in the hub triggered through gravity-driven onward transmission. Each hub j is assigned a color (the color of the surrounding box), and then all locations i are colored with intensity proportional to the probability $\mathbf{P}_{ij}$ that hub j sparked its outbreak (see Section 2). The prevailing black in California indicates that outbreaks in that state can be chiefly attributed to the hub in Stockton, CA. The purple in the eastern US indicates mixing of transmission chains seeded from Grenada, MS (blue) and Albany, GA (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

(Cooper et al., 2006; Colizza et al., 2007a), these results indicate that other critical ingredients are needed to explain the spatial introduction patterns of the autumn wave of the pandemic in the continental US.

One reason why the four observed transmission hubs do not correspond to major urban centers could be simply that the majority of the US population does not reside in cities. To illustrate this, it is necessary to shift attention away from ZIPs, which do not generally reflect an especially epidemiologically or socially relevant partition of the US population. The 2010 US Census' definition of an incorporated place (United States Census Bureau, 2015) corresponds more directly to the common notion of a city. There are 26 incorporated places in the US with population size greater than 600,000 (just over the size of the largest hub). These cities account for less than a fifth (15.5%) of the total US population. So, a given city-dweller would need over five-fold higher odds of sparking an epidemic than a non-city-dweller, for the probability of observing a transmission hub in a city to exceed that of observing a transmission hub elsewhere.

Previous immunity, school onsets, and meteorological effects may have tipped the balance further toward early outbreak establishment in these four hubs. An early wave of A/H1N1pdm influenza struck some major US cities, including New York and Chicago, between April and June of 2009, and may have conferred some immunity on those cities' populations. A protective effect from this underlying immunity could not be detected in previous analysis by Gog et al. (2014), and so was not considered here, but we cannot rule out the possibility that underlying immunity had some influence on the geographic placement of the hubs. The mixing of children in schools at the start of the autumn school term may also have increased the likelihood of epidemic establishment. Chao et al. (2010) provide evidence of this at the state level. Though the start of the autumn school term was not selected here as a key driver of short-distance influenza transmission (further discussion on this point to follow), it is possible that mixing in schools may have facilitated the establishment of long-distance jumps of infection. Importantly, the median school start dates in Grenada MS and Albany GA precede the school start date in Atlanta GA (the nearest large city) by one half week and one week, respectively. Furthermore, a cluster of six ZIPs surrounding and including Albany GA (and excluding Atlanta) had the country's earliest school start dates in the autumn of 2009. Though the difference in school term timing between this cluster and Atlanta is slight (about one-half to one week), this could explain why Albany, rather than Atlanta, was an epicenter of transmission for the eastern half of the US. More detailed data is needed to determine whether similar differences in school term timing are associated with the other hubs.

Meteorological factors such as humidity may have influenced the geography of the hubs. Ambient absolute humidity has been linked to the survival and subsequent transmissibility of the influenza virus (Lipsitch and Viboud, 2009; Shaman et al., 2011). Indeed, Shaman et al. (2011) correctly predicted a third pandemic wave in the southeastern US based on a spatiotemporal model of the effective reproductive number R_E driven by absolute humidity. The results presented here show that the southeast also played a crucial role in the spread of the second (autumn) pandemic wave, since the two most influential hubs lie in that region. This warrants further investigation of meteorological effects that may have predisposed the southeast to outbreak establishment in 2009.

It is impossible to identify or assess the importance of international hubs using the present dataset. This may especially affect inferences for the southwestern United States, since a major H1N1pdm outbreak was also occurring in the central and northern states of Mexico during the autumn of 2009 (Chowell et al., 2011). For example, the influenza activity in southern California, which is currently traced with high probability to the hub in Stockton (see Fig. 4), might be explained better by some unobserved hub just across the US-Mexico border. This issue highlights the need for fine-scale influenza incidence data that can be compared across national boundaries.

The transmission model considered here, Eq. (7), departs from the model developed by Gog et al. (2014) in three important ways. First, the present model is fit to data of a finer geographic resolution, the 3-dight ZIP, rather than the SCF. SCFs, like ZIPs, are designations made by the United States Postal Service, and consist of 2–3 3-digit ZIPs on average. To our knowledge, the analysis presented here represents the finest-scale mechanistic spatial model of influenza in the United States, though other investigations into different aspects of influenza transmission have considered data at a similar spatial scale (Yang et al., 2015a; Rumoro et al., 2014).

Second, we find that an exponential distance kernel captures the spatial dynamics of the autumn 2009 A/H1N1pdm influenza outbreak than a more traditional power kernel. In Gog et al. (2014), and in many gravity model-based descriptions of human mobility, a power kernel is used (Eggo et al., 2011; Xia et al., 2004; Mills and Riley, 2014). On the other hand, an exponential kernel is considered in Batty and Sikdar (1982), Gatto et al. (2012), and a few studies include some treatment of both (Liu et al., 2015; Ubøe, 2004; Truscott and Ferguson, 2012). In the present analysis, the exponential kernel is preferred heavily over the power kernel (see Table 1). It is difficult to justify *a priori* any one form of distance kernel over another. The preference of the exponential kernel is evidence that, on the whole, short-distance spread was the dominant mode of transmission of the 2009 A/H1N1pdm pandemic within the US. This reinforces the central role of short-range transmission during the 2009 pandemic, and as raised in Gog et al. (2014), this could be indicative of the importance of children as they have more localized mobility patterns. Data for the movement of children in the US is still lacking, but recent work in the UK suggests that school age children typically travel shorter distances each day than adults (Klepac et al., 2018). The strong preference for the exponential kernel observed here also provides evidence that power kernels should not necessarily be used as the 'null' assumption in spatial disease transmission models.

The third departure from Gog et al. (2014) regards the rejection of the start of the autumn school term as a predictor of short-distance transmission. At first glance, this is surprising, since schools in the US open in a south-to-north pattern (Fig. S27), much like the trajectory of the autumn 2009 pandemic wave, and since the outbreaks in the southeast coincided closely with the start of the autumn school term in that region (see Fig. S28). Empirical evidence also suggests that the mobility patterns of children may change significantly between term-time and vacation (Kucharski et al., 2015). However, the autumn 2009 pandemic wave lagged well behind the 'wavefront' of opening schools, such that the onset of the influenza outbreak in some locations in the northeast occurred up to eight weeks after the start of the school term in that region. So, the mixing of children in schools did not drive the spatial transmission of the influenza pandemic, so much as it provided 'fertile ground' for the pandemic to spread in its own time, which is not as easily detected by the model. Interestingly, if school start dates are advanced by one week, the result is a model with slightly better AIC than Eq. (7) (a 3-point improvement), that retains an effect from school start dates (see SI §4). This could provide evidence that mixing in children up to a week before the start of the autumn school term contributed to the spatial transmission of influenza. A virtually equivalent one-week shift was implicitly included in the transmission model presented by Gog et al. (2014), since the threshold outbreak onset detection method introduces an artificial 0.5- to 1-week bias toward later onsets (see Fig. S6). It is difficult to draw a clear conclusion on the role of schools using the present dataset and model, but fortunately, a clear conclusion need not be drawn here; the set of hubs is identified, with the same relative importances, whether using the transmission model Eq. (7) (without school start dates) or Eq. (S.6) (with one-week-advanced school start dates).

Geographic incidence data make it possible to identify pathogen establishment sites, as in this study and in Yang et al. (2015b). A complementary approach for inferring establishment sites uses genetic data instead, as in Lycett et al. (2012) and Lu et al. (2014). Linking

epidemiological and virological observations has proven difficult for human influenza (Viboud et al., 2013). The methods presented here may help bridge the gap by providing a spatially-detailed, testable hypothesis of the mixing patterns one might expect to see in spatially-referenced sequence data. First, one would need to test whether the phylogeographic patterns obtained from the influenza genomic data collected from 2009 (as from FluDB (National Institute of Allergy and Infectious Diseases, 2016) and GenBank (National Center for Biotechnology Information, 2016)) resemble the basins of infection depicted in Fig. 4. Then, the genomic data could be used to refine the reconstruction of between-city transmission chains, and to determine possible links between the long-distance jumps that led to epidemic establishment in the hubs. Combining the data streams in this way would shed more light on the true transmission network of the 2009 pandemic, improving in turn our ability to develop effective and efficient interventions for future outbreaks.

Author contributions

Conceived of the IMS data study and generated the database: LS, CV, BTG. Analyzed the data: SMK, JRG, CV. Developed methods: SMK, JRG, CV, VC, ONB, LS, BTG. Wrote the paper: SMK, JRG, CV, ONB, LS, BTG.

Funding

SMK was funded by a Gates Cambridge scholarship.

Competing interests

The authors have declared that no competing interests exist. LS acknowledges she consulted during 2007–2010 for SDI (which later became part of IMS).

Acknowledgments

Data for this article were made available by SDI Health (now Quintiles IMS) by agreement with the RAPIDD program of the Science and Technology Directorate, Department of Homeland Security. Conversations at a September 2015 workshop hosted by the RAPIDD program contributed to the conceptualization of this work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.epidem.2018.10.002.

References

Abeku, T.A., Hay, S.I., Ochola, S., Langi, P., Beard, B., de Vlas, S.J., Cox, J., 2004. Malaria epidemic early warning and detection in African highlands. Trends Parasitol. 20 (9), 400–405.
Batty, M., Sikdar, P.K., 1982. Spatial aggregation in gravity models: 2. One-dimensional population density models. Environ. Plan. A 14 (4), 525–553.
Burnham, K.P., Anderson, D.R., 2004. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer-Verlag, New York.
Carneiro, H.A., Mylonakis, E., 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin. Infect. Dis. 49 (10), 1557–1564.
Centers for Disease Control and Prevention, 2016. Overview of Influenza Surveillance in the United States. Technical Report. Centers for Disease Control and Prevention.
Chao, D.L., Halloran, M.E., Longini, I.M., 2010. School opening dates predict pandemic influenza A(H1N1) outbreaks in the United States. J. Infect. Dis. 202 (6), 877–880.
Charu, V., Zeger, S., Gog, J., Bjørnstad, O.N., Kissler, S., Simonsen, L., Grenfell, B.T., Viboud, C., 2017. Human mobility and the spatial transmission of influenza in the United States. PLoS Comput. Biol. 13 (2), e1005382.
Chowell, G., Echevarría-Zuno, S., Viboud, C., Simonsen, L., Tamerius, J., Miller, M.A., Borja-AburtoVí, V.H., 2011. Characterizing the epidemiology of the 2009 influenza A/H1N1 pandemic in Mexico. PLoS Med. 8 (5).
Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.J., Vespignani, A., 2007a. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med. 4 (1), 0095–0110.
Colizza, V., Barthélemy, M., Barrat, A., Vespignani, A., 2007b. Epidemic modeling in complex realities. Comptes Rendus Biol. 330 (4), 364–374.
Cooper, B.S., Pitman, R.J., Edmunds, W.J., Gay, N.J., 2006. Delaying the international spread of pandemic influenza. PLoS Med. 3 (6), 0845–0855.
Eggo, R.M., Cauchemez, S., Ferguson, N.M., 2011. Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. J. R. Soc. Interface 8 (55), 233–243.
Ferguson, N.M., Cummings, D.A.T., Fraser, C., Cajka, J.C., Cooley, P.C., Burke, D.S., 2006. Strategies for mitigating an influenza pandemic. Nature 442 (7101), 448–452.
Galvani, A.P., May, R.M., 2005. Epidemiology: dimensions of superspreading. Nature 438 (7066), 293–295.
Gatto, M., Mari, L., Bertuzzo, E., Casagrandi, R., Righetto, L., Rodriguez-Iturbe, I., Rinaldo, A., 2012. Generalized reproduction numbers and the prediction of patterns in waterborne disease. Proc. Natl. Acad. Sci. U. S. A. 109 (48), 19703–19708.
Gog, J.R., Ballesteros, S., Viboud, C., Simonsen, L., Bjørnstad, O.N., Shaman, J., Chao, D.L., Khan, F., Grenfell, B.T., 2014. Spatial transmission of 2009 pandemic influenza in the US. PLoS Comput. Biol. 10 (6), e1003635.
Grad, Y.H., Lipsitch, M., 2014. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. Genome Biol. 15 (11), 538.
Hashimoto, S., Murakami, Y., Taniguchi, K., Nagai, M., 2000. Detection of epidemics in their early stage through infectious disease surveillance. Int. J. Epidemiol. 29 (5), 905–910.
Held, L., Hofmann, M., Höhle, M., Schmid, V., 2006. A two-component model for counts of infectious diseases. Biostatistics 7 (3), 422–437.
Kilpatrick, A.M., Chmura, A.A., Gibbons, D.W., Fleischer, R.C., Marra, P.P., Daszak, P., 2006. Predicting the global spread of H5N1 avian influenza. Proc. Natl. Acad. Sci. U. S. A. 103 (51), 19368–19373.
Klepac, P., Kissler, S., Gog, J., 2018. Contagion! The BBC Four Pandemic – the model behind the documentary. Epidemics 24, 49–59.
Kucharski, A.J., Conlan, A.J., Eames, K.T., 2015. School's out: seasonal variation in the movement patterns of school children. PLOS ONE 10 (6), 1–10.
Lampos, V., Bie, T.D., Cristianini, N., 2010. Flu detector – tracking epidemics on Twitter. In: Balcazar, J., Bonchi, F., Sebag, M., Gionis, A. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, Berlin/Heidelberg, pp. 599–602.
Legrand, J., Egan, J.R., Hall, I.M., Cauchemez, S., Leach, S., Ferguson, N.M., 2009. Estimating the location and spatial extent of a covert anthrax release. PLoS Comput. Biol. 5 (1).
Levy, M.Z., Small, D.S., Vilhena, D.A., Bowman, N.M., Kawai, V., Cornejo del Carpio, J.G., Cordova-Benzaquen, E., Gilman, R.H., Bern, C., Plotkin, J.B., 2011. Retracing micro-epidemics of Chagas disease using epicenter regression. PLoS Comput. Biol. 7 (9).
Lipsitch, M., Viboud, C., 2009. Influenza seasonality: lifting the fog. Proc. Natl. Acad. Sci. U. S. A. 106 (10), 3645–3646.
Liu, H., Chen, Y.-H., Lih, J.-S., 2015. Crossover from exponential to power-law scaling for human mobility pattern in urban, suburban and rural areas. Eur. Phys. J. B 88 (5), 117.
Lu, L., Lycett, S.J., Leigh Brown, A.J., 2014. Determining the phylogenetic and phylogeographic origin of highly pathogenic avian influenza (H7N3) in Mexico. PLOS ONE 9 (9), e107330.
Lycett, S., Mcleish, N.J., Robertson, C., Carman, W., Baillie, G., McMenamin, J., Rambaut, A., Simmonds, P., Woolhouse, M., Leigh Brown, A.J., 2012. Origin and fate of A/H1N1 influenza in Scotland during 2009. J. Gen. Virol. 93 (Pt 6), 1253–1260.
Mills, H.L., Riley, S., 2014. The spatial resolution of epidemic peaks. PLoS Comput. Biol. 10 (4) e1003561.
Moriyama, I.M., Loy, R.M., Robb-Smith, A.H., 2011. History of the Statistical Classification of Diseases and Causes of Death. Technical Report. Centers for Disease Control and Prevention.
National Center for Biotechnology Information, 2016. Influenza Virus Resource.
National Institute of Allergy and Infectious Diseases, 2016. Influenza Research Database.
Olson, D.R., Konty, K.J., Paladini, M., Viboud, C., Simonsen, L., 2013. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput. Biol. 9 (10), e1003256.
Paull, S.H., Song, S., McClure, K.M., Sackett, L.C., Kilpatrick, A.M., Johnson, P.T., 2012. From superspreaders to disease hotspots: linking transmission across hosts and space. Front. Ecol. Environ. 10 (2), 75–82.
Pelecanos, A.M., Ryan, P.A., Gatton, M.L., 2010. Outbreak detection algorithms for seasonal disease data: a case study using Ross River virus disease. BMC Med. Inform. Decis. Mak. 10 (1), 74.
Riley, S., Eames, K., Isham, V., Mollison, D., Trapman, P., 2015. Five challenges for spatial epidemic models. Epidemics 10, 68–71.
Rumoro, D., Shah, S., Trenholme, G., Gibbs, G., Hallock, M., Waddell, M.J., 2014. Creating a local geographic influenza-like illness activity report. ISDS Annual Conference Proceedings 2014, vol. 7 2579.
Russell, C.A., Jones, T.C., Barr, I.G., Cox, N.J., Garten, R.J., Gregory, V., Gust, I.D., Hampson, A.W., Hay, A.J., Hurt, A.C., de Jong, J.C., Kelso, A., Klimov, A.I., Kageyama, T., Komadina, N., Lapedes, A.S., Lin, Y.P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A.D.M.E., Rimmelzwaan, G.F., Shaw, M.W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R.A.M., Smith, D.J., 2008. The global circulation of seasonal influenza A (H3N2) viruses. Science 320 (5874), 340–346.
Shaman, J., Goldstein, E., Lipsitch, M., 2011. Absolute humidity and pandemic versus epidemic influenza. Am. J. Epidemiol. 173 (2), 127–135.
Shmueli, G., Burkom, H., 2010. Statistical challenges facing early outbreak detection in biosurveillance. Technometrics 52 (1), 39–51.
Skene, K.J., Paltiel, A.D., Shim, E., Galvani, A.P., 2014. A marginal benefit approach for vaccinating influenza "superspreaders". Med. Decis. Mak. 34 (4), 536–549.
Tatem, A.J., 2014. Mapping population and pathogen movements. Int. Health 6 (1), 5–11.

ARTICLE IN PRESS

Truscott, J., Ferguson, N.M., 2012. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. PLoS Comput. Biol. 8 (10), e1002699.
Ubøe, J., 2004. Aggregation of gravity models for journeys to work. Environ. Plan. A 36 (4), 715–729.
United States Census Bureau, 2015. U.S. Gazetteer Files.
United States Postal Service, 2018. 3-Digit ZIP Code Prefix Matrix.
U.S. Postal Service Office of Inspector General, 2013. The Untold Story of the ZIP Code. Technical report. United State Postal Services.
Viboud, C., Charu, V., Olson, D., Ballesteros, S., Gog, J., Khan, F., Grenfell, B., Simonsen, L., 2014. Demonstrating the use of high-volume electronic medical claims data to monitor local and regional influenza activity in the US. PLOS ONE 9 (7), e102429.
Viboud, C., Nelson, M.I., Tan, Y., Holmes, E.C., 2013. Contrasting the epidemiological and evolutionary dynamics of influenza spatial transmission. Philos. Trans. R. Soc. B: Biol. Sci. 368 (1614), 20120199.
Wagner, M.M., Tsui, F.C., Espino, J.U., Dato, V.M., Sittig, D.F., Caruana, R.a., McGinnis, L.F., Deerfield, D.W., Druzdzel, M.J., Fridsma, D.B., 2001. The emerging science of very early detection of disease outbreaks. J. Public Health Manag. Pract. 7 (290), 51–59.
WHO Global Influenza Programme Surveillance and Epidemiology team, July 2012. WHO Interim Global Epidemiological Surveillance Standards for Influenza. Technical Report. World Health Organization.
Wilson, A., 1970. Entropy in Urban and Regional Modelling. Pion, London.
Won, M., Marques-Pita, M., Louro, C., Gonçalves-Sá, J., Barker, W.H., Molinari, N.-A.M., Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L., Hickmann, K., Fairchild, G., Priedhorsky, R., Generous, N., Hyman, J., Deshpande, A., Santillana, M., Nguyen, A.T., Dredze, M., Paul, M.J., Nsoesie, E.O., Brownstein, J.S., Chretien, J., George, D., Shaman, J., Chitale, R., McKenzie, F., Shaman, J., Alicia, K., Moriña, D., Rhodes, C.J., Hollingsworth, T.D., Closas, P., Coma, E., Méndez, L., Cowling, B.J., Martinez-Beneito, M.A., Shaman, J., Christakis, N.A., Fowler, J.H., Pervaiz, F., Kermack, W., McKendrick, A., Trevor, J.H., Tibshirani, R.J., Friedman, Lazer, D., Kennedy, R., King, G., Vespignani, A., Olson, D.R., Konty, K.J., Paladini, M., Viboud, C., Simonsen, L., Tariq, A., Westbrook, J., Byrne, M., Robinson, M., Baysari, M.T., Cooper, D.L., 2017. Early and real-time detection of seasonal influenza onset. PLoS Comput. Biol. 13 (2), e1005330.
Xia, Y., Bjørnstad, O.N., Grenfell, B.T., 2004. Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. Am. Nat. 164 (2), 267–281.
Yang, W., Lipsitch, M., Shaman, J., 2015a. Inference of seasonal and pandemic influenza transmission dynamics. Proc. Natl. Acad. Sci. U. S. A. 112 (9), 201415012.
Yang, W., Zhang, W., Kargbo, D., Yang, R., Chen, Y., Chen, Z., Kamara, A., Kargbo, B., Kandula, S., Karspeck, A., Liu, C., Shaman, J., 2015b. Transmission network of the 2014–2015 Ebola epidemic in Sierra Leone. J. R. Soc. Interface 12 (112), 20150536.
Zipf, G.K., 1946. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. Am. Sociol. Rev. 11 (6), 677.