Ming *et al. Infectious Diseases of Poverty* (2016) 5:107
DOI 10.1186/s40249-016-0199-5

Infectious Diseases of Poverty

RESEARCH ARTICLE Open Access

Stochastic modelling of infectious diseases for heterogeneous populations

Rui-Xing Ming[1], Jiming Liu[2], William K. W. Cheung[2] and Xiang Wan[3*]

Abstract

Background: Infectious diseases such as SARS and H1N1 can significantly impact people's lives and cause severe social and economic damages. Recent outbreaks have stressed the urgency of effective research on the dynamics of infectious disease spread. However, it is difficult to predict when and where outbreaks may emerge and how infectious diseases spread because many factors affect their transmission, and some of them may be unknown.

Methods: One feasible means to promptly detect an outbreak and track the progress of disease spread is to implement surveillance systems in regional or national health and medical centres. The accumulated surveillance data, including temporal, spatial, clinical, and demographic information can provide valuable information that can be exploited to better understand and model the dynamics of infectious disease spread. The aim of this work is to develop and empirically evaluate a stochastic model that allows the investigation of transmission patterns of infectious diseases in heterogeneous populations.

Results: We test the proposed model on simulation data and apply it to the surveillance data from the 2009 H1N1 pandemic in Hong Kong. In the simulation experiment, our model achieves high accuracy in parameter estimation (less than 10.0 % mean absolute percentage error). In terms of the forward prediction of case incidence, the mean absolute percentage errors are 17.3 % for the simulation experiment and 20.0 % for the experiment on the real surveillance data.

Conclusion: We propose a stochastic model to study the dynamics of infectious disease spread in heterogeneous populations from temporal-spatial surveillance data. The proposed model is evaluated using both simulated data and the real data from the 2009 H1N1 epidemic in Hong Kong and achieves acceptable prediction accuracy. We believe that our model can provide valuable insights for public health authorities to predict the effect of disease spread and analyse its underlying factors and to guide new control efforts.

Keywords: Epidemiology, Stochastic model, Surveillance system, Spread pattern

Multilingual abstracts

Please see Additional file 1 for translations of the abstract into the five official working languages of the United Nations.

Background

Infectious diseases remain a major cause of morbidity and mortality worldwide, triggering immeasurable loss in many societies. Most people may still have a fresh memory of the H1N1 outbreak in 2009, which brought pictures of empty streets and people wearing face masks and collectively caused at least 12799 deaths according to the World Health Organization (WHO) report [1]. The H1N1 pandemic calls for research on accurately modelling the spread dynamics of an infectious disease, which offers a practically useful means for policy makers to evaluate the potential effects of intervention strategies [2–4].

Mathematical models of the spread of infectious diseases are an important tool for investigating and quantifying the spread dynamics because direct experimental study on the spread of disease among humans is not ethical. Although the subjects involved in different epidemics may be different, many can be modeled by the popular Susceptible-Infected-Recovered (SIR) models [5–7], which study the spread of infectious diseases by

*Correspondence: xwan@comp.hkbu.edu.hk
[3]Department of Computer Science & Institute of Theoretical and Computational Study, Hong Kong Baptist University, Kowloon Tong, Hong Kong
Full list of author information is available at the end of the article

tracking the number (S) of people susceptible to the disease, the number (I) of people infected with the disease, and the number (R) of people who have recovered from the disease. Three assumptions are made: (1) the total population $N = S(t) + I(t) + R(t)$ is fixed at any time t; (2) those who have recovered from the disease are forever immune; and (3) those who have not had the disease are equally susceptible, and the probability of their contracting the disease at time t is proportional to the product of $S(t)$ and $I(t)$. Based on these assumptions, the SIR model defines a set of three ordinary differential equations for S(t), I(t), and R(t):

$$\begin{aligned} dS/dt &= -\beta S(t)I(t) \\ dI/dt &= \beta S(t)I(t) - kI(t) \\ dR/dt &= kI(t). \end{aligned} \quad (1)$$

Here, $\beta \geq 0$ is the effective transmission rate and $k \geq 0$ is the recovery rate. Because the SIR-based models are well presented in the literature, herein, we omit a verbose introduction of these models. Readers with an interest in such a topic can find the details in [5–7].

The SIR-based models and its variants have proven to be quite useful in the study of the spread dynamics of infectious diseases [8–10]. In [11–13], the progression of disease spread is characterized by tracking the number of S_t with a chain binomial model. The number of susceptible members $S_{t+\triangle t}$ ($\triangle t$ represents the infectious period of the disease and is always chosen to be $1/k$) at time $t + \triangle t$ is a binomial random variable that depends on S_t and $I_t\alpha$, $S_{t+\triangle t} \sim Bin(S_t, 1 - I_t\alpha)$, which provides a recursive relationship between $S_{t+\triangle t}$ and S_t and produces a formal stochastic process. However, the power of these models is mainly limited to uniform and homogeneous populations or populations with infinite size and homogeneous interactions. In many cases, the actual spread of infectious diseases occurs in a diverse or dispersed population. To study the spread of infectious diseases in heterogeneous populations, people usually divide a population into subpopulations that differ from each other. Sub-populations can be determined on the basis of social, cultural, economic, demographic, and geographic factors. Next, besides the dynamics of the internal spread within a subpopulation, the transmission dynamics between subpopulations should also be considered in the study of epidemic spreading.

Network-based epidemic modelling represents a popular approach for heterogeneous populations in which the nodes in the network correspond to sub-populations, and the links indicate the neighboring relationships. Many network-based models have been proposed, including patch models [14–16], distance-transmission models [17], and multi-group models [18, 19]. However, these models require knowledge of every individual (or host) and all relationships between individuals, which may be not achievable due to information privacy-related restrictions and the high cost of subject recruitment. To overcome the difficulties of collecting data, researchers have investigated several types of computer-generated networks in the context of disease spread in population-scale studies [20–24]. Grassberger first studied the dynamics of infectious diseases that propagate on regular networks using the percolation theory [25]. Recent studies have revealed that many real-world networks, including social networks in which infectious diseases propagate, are either small-world [26] or scale-free [27] rather than regular or random, as thought previously [28]. Because the underlying structures of networks will influence the effect that the dynamics of epidemics will have on them, researchers, such as Pastor-Satorras and Vespignani, have made many contributions to critical value analysis of typical epidemics on different types of complex network [23, 24, 29]. On the basis of the mean-field theory, they found that compared with homogeneous networks, scale-free networks are fragile to the invasion of infectious diseases, computer viruses, or any other type of negative epidemics.

Epidemics have also been studied in various disciplines. Sociologists are concerned with the diffusion of rumors or innovation on social networks [30]; economists have studied viral marketing and recommendation strategies by considering both cascading dynamics and the network effects of vital nodes [31]; and computer scientists are interested in how some topics can quickly cascade in virtual blog spaces and how their propagation trends [32, 33].

Although network-based studies have contributed to the modelling of disease and/or information dynamics, some models make a strong assumption that the structures of underlying networks over which epidemics spread are known beforehand. In the real world, however, the structures of underlying diffusion networks are not known directly. Many others assume the availability of information about the interactions occurring between individuals [34–37] that are often not valid in the context of disease spread. What may be obtained is only the time at which particular sub-populations become infected, but not how they become infected, nor how they affect their neighboring areas. Moreover, the underlying structures of networks will greatly influence the dynamics of infectious disease spread.

Since the emergence of the H1N1 influenza pandemic in April 2009, its underlying dynamics have been of great public health interest, and many approaches for its study have been proposed [14, 38–41]. Most of them are based on the classic SIR model. For example, Birrell et al. [40] provided an age structure-based compartmental model with a Bayesian synthesis of multiple evidence sources to reveal substantial changes in contact patterns throughout

the epidemic. Besides of the compartmental models, other mathematical models are also used to describe the transmission dynamics [3, 42–47]. The chain binomial model was used to calculate the household secondary attack rates to measure the transmissibility of the 2009 H1N1 influenza pandemic by Lessler et al. [44] and Klick et al. [45]. Yang et al. [46] constructed a model based on chains of infections and used the infection hazard function and survival function to study the 2009 H1N1 influenza pandemic. Ferguson et al. [3] and Cauchemez et al. [42, 43] incorporated other factors, such as household risk, within-school risk, and community risk, in the study of infection spread and found out that younger age groups under 19 years old were more susceptible than older age groups. Jin et al. [47] formulated an epidemic model of influenza A based on networks and calculated the basic reproduction number and studied the effects of various immunization schemes. However, this work required that the individual contact pattern be provided. Nonetheless, none of the aforementioned approaches takes spatial heterogeneity into consideration in the study of disease spread.

Recently, an outbreak of Ebola virus disease (EVD) swept across parts of West Africa from March 2014 to April 2015. By June 10, 2015, WHO had reported 27,237 confirmed, probable, or suspected cases in three countries with 11,158 deaths [48]. This epidemic received extensive research attention on its dynamics of spread [49–57] (for further references in the review article [58]). To name a few, Chowell et al. found that district-level Ebola virus disease outbreaks in West Africa follow polynomial-based growth in time instead of the exponential growth that describes the progress of many infectious disease epidemics [52]. Fisman et al. used a simple, two parameter mathematical model to characterize epidemic growth patterns in the 2014 Ebola outbreak [53]. Webb et al. proposed a variant of the classic SIR model with three extra groups, incubating, contaminated and isolated, which can provide a more accurate prediction for the future incidences [56]. Carroll et al. used a deep sequencing approach to gain insight into the evolution of the Ebola virus (EBOV) in Guinea from the ongoing West African outbreak. The viral sequence data can be combined with epidemiological information to retrospectively test the effectiveness of control measures, and provides an unprecedented window into the evolution of an ongoing outbreak of viral haemorrhagic fever [57].

To accurately predict when and where outbreaks will occur, a feasible means is to deploy manual or electronic surveillance systems through regional or national public health and medical organizations [59]. Most of the surveillance data accumulated from such systems contains temporal, spatial, clinical, and demographic information. For instance, Telehealth Ontario is a teletriage helpline that is available free to all Ontario residents, which allows those with suspected infections to connect with experts who can assess their symptoms. The records of such calls provide valuable information on which individual from where was possibly infected and by which type of disease at what time. In this paper, we address the problem of modelling disease spread dynamics in heterogeneous populations from temporal-spatial surveillance data. We analyse the role of heterogeneity in a stochastic epidemic model on a two-dimensional lattice. Within a particular sub-population, the speed of spread is controlled by a single parameter, the transmissibility of the pathogen between individuals. Between sub-populations, the transmissibility becomes a random variable drawn from a probability distribution. Our work differs from existing studies in some fundamental ways, in light of the unique nature of infectious disease diffusion dynamics. Our results have practical implications for the analysis of disease control strategies in realistic heterogeneous epidemic systems.

Methods

In this work, we propose a stochastic model to study the dynamics of infectious disease spread in heterogeneous populations from temporal-spatial surveillance data. We divide the whole population into m sub-populations on the basis of geographic regions. In the following, we use $S_i(t), I_i(t)$, and $R_i(t)$ to denote the number of susceptible, infected, and recovered people, respectively, at time t in region i, $i = 1, 2, \cdots, m$ and $t \in [0, T]$.

Stochastic model

Classic SIR-based modelling of infectious diseases assumes that the population is well-mixed. To take the role of heterogeneity into consideration, we use an alternative approach to model the dynamics of infectious disease spread. First, the classic SIR model (Eq. (1)) studies the change in the numbers of peoples in the three groups. In reality, the change in the number of the infected people is the major concern of society. Second, in many epidemics or pandemics such as H1N1 and SARS, the number of infected people $I_i(t)$ is relatively small compared to the whole subpopulation $S_i(t)$. Therefore, we may consider $S_i(t)$ as a constant to simplify the modelling of the change in the number of infected people $I(t)$, for which we propose the following stochastic differential equation:

$$\mathrm{d}I_i(t) = (\alpha + \delta_i I_i(t))\mathrm{d}t + \sigma_i \mathrm{d}B_i(t), \tag{2}$$

where α is a parameter that measures the auto-recovery rate of one particular infectious disease, which is usually considered as a constant among sub-populations, δ_i is the parameter that measures the different disease transmissibility in different subpopulations, $\sigma_i > 0$ is the diffusion

parameter that measures the disease spread from neighbors, and $B_i(t)$ is a standard Brownian motion. It is worth noting that we assume the parameter $\delta_i \neq 0$ for technical purposes, and the results in the case of $\delta_i = 0$ can be achieved with $\delta_i \to 0$.

Comparing our model in Eq. (2) with the classic model in Eq. (1), we can see that they both capture the situation in which the change in the number of infected people has a positive relationship with the total number of infected people, which means that the more infected people there are, the more people will get infected. There are two key differences between these two models: first, the key factor $(\beta S_i(t) - k)$ associated with the disease spread in Eq. (1) is replaced with a single parameter δ_i in Eq. (2), which can be used to analyse the role of heterogeneity in the disease spread; and second, Eq. (2) takes the neighboring relationships into consideration to study the dynamics of the disease spread among different sub-populations.

By Ito formula, the solution of Eq. (2) is given by

$$\begin{aligned} I_i(t) &= I_i(0)e^{\delta_i t} + \tfrac{\alpha}{\delta_i}(e^{\delta_i t} - 1) \\ &\quad + \sigma_i e^{\delta_i t}\textstyle\int_0^t e^{-\delta_i s}\mathrm{d}B_i(s). \end{aligned} \tag{3}$$

Notice that for any fixed t, $\int_0^t e^{-\delta_i s}\mathrm{d}B_i(s)$ is a normal random variable with

$$\begin{aligned} E\left[\int_0^t e^{-\delta_i s}\mathrm{d}B_i(s)\right] &= 0, \\ Var\left[\int_0^t e^{-\delta_i s}\mathrm{d}B_i(s)\right] &= \frac{1 - e^{-2\delta_i t}}{2\delta_i}. \end{aligned} \tag{4}$$

Thus, for any fixed t, $I_i(t)$ is a normal random variable with

$$E[\,I_i(t)] = I_i(0)e^{\delta_i t} + \frac{\alpha}{\delta_i}\left(e^{\delta_i t} - 1\right) \tag{5}$$

and

$$Var[\,I_i(t)] = \frac{\sigma_i^2}{2\delta_i}\left(e^{2\delta_i t} - 1\right). \tag{6}$$

There are three cases of being interested for parameter α:

- $\alpha > -I_i(0)\delta_i$
  In this case, $E[\,I_i(t)]$ tends to infinity as t goes to infinity, which implies that all people in that region will be infected if the time is long enough.
- $\alpha = -I_i(0)\delta_i$
  In this case, the pandemic or epidemic will reach a state of equilibrium.
- $\alpha < -I_i(0)\delta_i$
  In this case, $E[\,I_i(t)]$ will reach 0 at some time $t = \hat{t}$ and go to negative infinity as t goes to infinity, which implies the pandemic or epidemic will end at time $\hat{t}$.

Parameter estimation

To estimate the parameters in our proposed stochastic model from the surveillance data, we need to divide the interval $[\,0, T]$ into n subintervals, $[\,t_0, t_1]$, $[\,t_1, t_2]$, $\cdots$, $[\,t_{n-1}, t_n]$, where $0 = t_0 < t_1 < t_2 < \cdots < t_n = T$. Denote $\triangle t(k) = t_{k+1} - t_k$, $\triangle B_i(k) = B_i(t_{k+1}) - B_i(t_k)$, $\triangle I_i(k) = I_i(t_{k+1}) - I_i(t_k)$, $k = 0, 1, \cdots, n-1$. Then Eq. (2) is rewritten as

$$\triangle I_i(k) = (\alpha + \delta_i I_i(t_k))\triangle t(k) + \sigma_i \triangle B_i(k). \tag{7}$$

It is easy to see that $\triangle I_i(k)|I_i(t_k) \sim N\big((\alpha + \delta_i I_i(t_k))\triangle t(k), \sigma_i^2 \triangle t(k)\big)$. Let $\theta_i = (\alpha, \delta_i, \sigma_i)$. Then the transition density of the process $\{I_i(t); t \geq 0\}$ is

$$\begin{aligned} &p_{\theta_i}(s + t, y|s, x) \\ &= \frac{1}{\sqrt{2\pi t\sigma_i^2}} \exp\left\{-\frac{(y - x - (\alpha + \delta_i x)t)^2}{2t\sigma_i^2}\right\}. \end{aligned} \tag{8}$$

Hence, the likelihood function is given by

$$\begin{aligned} &f(\theta_i|I_i) \triangleq f\,(\theta_i|I_i(t_k), 0 \leq k \leq n-1) \\ &= I_i(0)\left(\frac{1}{2\pi\sigma_i^2}\right)^{\frac{n}{2}} \prod_{k=0}^{n-1} \frac{1}{\triangle t(k)} \\ &\quad \exp\left\{-\frac{(\triangle I_i(k) - (\alpha + \delta_i I_i(t_k))\triangle t(k))^2}{2\triangle t(k)\sigma_i^2}\right\}. \end{aligned} \tag{9}$$

Consequently, the log-likelihood function is

$$\begin{aligned} \log f(\theta_i|I_i) &\propto -\frac{n}{2}\log\sigma_i^2 \\ &\quad - \sum_{k=0}^{n-1} \frac{(\triangle I_i(k) - (\alpha + \delta_i I_i(t_k))\triangle t(k))^2}{2\triangle t(k)\sigma_i^2}. \end{aligned} \tag{10}$$

Let

$$u_{i1} = \frac{1}{T}\sum_{k=0}^{n-1} I_i(t_k), \tag{11}$$

$$u_{i2} = \frac{1}{T}\sum_{k=0}^{n-1} I_i(t_{k+1}), \tag{12}$$

$$u_{i11} = \frac{1}{T}\sum_{k=0}^{n-1} I_i^2(t_k), \tag{13}$$

$$u_{i12} = \frac{1}{T}\sum_{k=0}^{n-1} I_i(t_k)I_i(t_{k+1}), \tag{14}$$

$$u_{i1\Delta} = \frac{1}{T}\sum_{k=0}^{n-1} I_i(t_k)\triangle t(k), \tag{15}$$

$$u_{i11\Delta} = \frac{1}{T}\sum_{k=0}^{n-1} I_i^2(t_k)\triangle t(k), \tag{16}$$

$$u_{i11\Delta^{-1}} = \frac{1}{T}\sum_{k=0}^{n-1} I_i^2(t_k)(\triangle t(k))^{-1}, \quad (17)$$

$$u_{i12\Delta^{-1}} = \frac{1}{T}\sum_{k=0}^{n-1} I_i(t_k)I_i(t_{k+1})(\triangle t(k))^{-1}, \quad (18)$$

$$u_{i22\Delta^{-1}} = \frac{1}{T}\sum_{k=0}^{n-1} I_i^2(t_{k+1})(\triangle t(k))^{-1}. \quad (19)$$

We have the estimator of θ_i as follows:

$$\widehat{\delta_i} = \frac{u_{i12} - u_{i11} - u_{i2}u_{i1\Delta} + u_{i1}u_{i1\Delta}}{u_{i11\Delta} - u_{i1\Delta}^2}, \quad (20)$$

$$\widehat{\alpha} = u_{i2} - u_{i1} - \widehat{\delta_i}u_{i1\Delta}, \quad (21)$$

$$\widehat{\sigma_i^2} = Tn^{-1}\left\{u_{i22\Delta^{-1}} - 2u_{i12\Delta^{-1}} + u_{i11\Delta^{-1}} - (u_{i2} - u_{i1})^2 + [(u_{i2} - u_{i1})u_{i1\Delta} - (u_{i12} - u_{i11})]\widehat{\delta_i}\right\}. \quad (22)$$

It is obviously to see that $\widehat{\alpha}$ is not a bona fide estimator of α, because only the information of $\{I_i(t); 0 \le t \le T\}$ is used to estimate α. A good estimator should pool all the information $\{I_i(t); 0 \le t \le T\}$ $(i = 1, 2, \cdots, m)$. There are two ways to find the pool estimator. The first way is to approximate α by pooling all $\widehat{\alpha_i}$ as follows:

$$\widehat{\alpha} = m^{-1}\sum_{i=1}^{m}\widehat{\alpha_i}, \quad (23)$$

$$\widehat{\alpha_i} = u_{i2} - u_{i1} - \widehat{\delta_i}u_{i1\Delta}. \quad (24)$$

But the issue in Eq. (23) is that m must be very large in order to achieve the accurate estimate of α. In this work, we choose the second way, which is the maximum likelihood estimation. To do so, we need to assume that the processes $\{I_i(t); 0 \le t \le T\}$ $(i = 1, 2, \cdots, m)$ are mutually independent. Then the log-likelihood function of $\{I_i(t); 0 \le t \le T\}$ $(i = 1, 2, \cdots, m)$ is given by

$$\sum_{i=1}^{m}\log f(\alpha|I_i) \propto -\sum_{i=1}^{m}\sum_{k=0}^{n-1}\frac{(\triangle I_i(k) - (\alpha + \widehat{\delta_i}I_i(t_k))\triangle t(k))^2}{2\triangle t(k)\widehat{\sigma_i^2}}. \quad (25)$$

The maximum likelihood estimate is

$$\widetilde{\alpha} = \sum_{i=1}^{m}\omega_i\widehat{\alpha_i}, \quad (26)$$

where

$$\omega_j = \frac{\widehat{\sigma_j^2}^{-1}}{\sum_{i=1}^{m}\sum_{k=0}^{n-1}\widehat{\sigma_i^2}^{-1}}, \; j = 1, 2, \cdots, m. \quad (27)$$

$\widehat{\alpha_i}$ is defined in Eq. (24).

Results and discussion

In this section, we illustrate the performance of our proposed model using both simulated and real data.

Simulation study

In the simulation study, we examine the performance of our proposed model with respect to the accuracy of parameter estimation and the forward prediction of the case incidence. First, we generate data using various parameters by the following steps:

1. Set $m = 4$ (the number of sub-populations) and $T = 100$ (the number of time slots). These two numbers are randomly selected.
2. Randomly draw α from $[0.05, 0.09]$, δ_i from $[0.02, 0.08]$, and σ_i from $[0.02, 0.08]$.
3. Initialize $I_i(0), 1 \le i \le m$.
4. Simulate $I_i(k+1) = I_i(k) + \triangle I_i(k), k = 0, 1, \cdots, T-1$ using Eq. (7) and $\triangle I_i(k)|I_i(t_k) \sim N\left((\alpha + \delta_i I_i(t_k))\triangle t(k), \sigma_i^2\triangle t(k)\right)$.

Three parameters, α, δ_i, and σ_i, in Eq. (2) will be estimated from the simulated data. We conduct 100 replicates by repeating Step 2–4 and compare the estimated ones, $\hat{\alpha}, \hat{\delta_i}$, and $\hat{\sigma_i}$, with the ground truth values in terms of the mean absolute percentage error (MAPE) defined as:

$$E_\alpha = \frac{1}{100}\sum_{j=1}^{100}\left|\frac{\hat{\alpha}_j - \alpha_j}{\alpha_j}\right|, \quad (28)$$

$$E_\delta = \frac{1}{100 * m}\sum_{i=0}^{m-1}\sum_{j=1}^{100}\left|\frac{\hat{\delta}_{ij} - \delta_{ij}}{\delta_{ij}}\right|, \quad (29)$$

$$E_\sigma = \frac{1}{100 * m}\sum_{i=0}^{m-1}\sum_{j=1}^{100}\left|\frac{\hat{\sigma}_{ij} - \sigma_{ij}}{\sigma_{ij}}\right|. \quad (30)$$

The mean absolute percentage errors (MAPEs) for E_α, E_δ, and E_σ are 10.0 %, 6.0 %, and 10.0 %, respectively. We plot the distribution of the estimated errors for 100 replicates for $\hat{\alpha}, \hat{\delta_i}$, and $\hat{\sigma_i}$ in Fig. 1. From Fig. 1, we can see that both the estimates of $\hat{\alpha}$ and $\hat{\sigma_i}$ have small variations. The variation of the estimate of $\hat{\delta_i}$ is slightly larger but is still acceptable and is due to the uncertainty embedded in the stochastic process. We also use the estimated values of the parameters to generate the data and compare it with the simulated data using the ground truth values of the parameters. The correlation between them is 0.96. We randomly select one replicate and show the comparison results in Fig. 2. Basically, we can use the estimated parameters to accurately recover the ground truth data.

Next, we conduct an experiment to test the prediction accuracy of our model. Let us consider a sequence of data

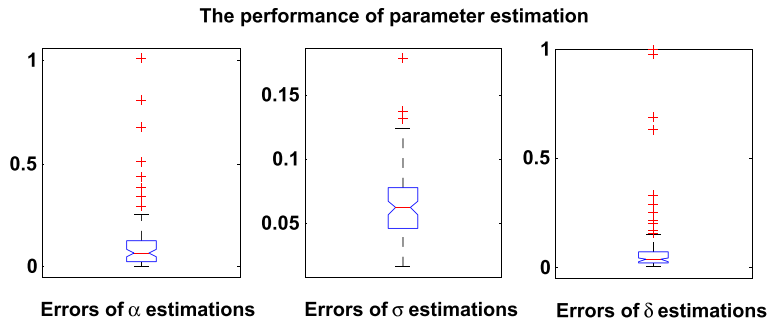

Fig. 1 The performance of parameter estimation. The mean absolute percentage errors for α, δ, and σ are 10.0, 6.0, and 10.0 %, respectively. α measures the auto-recovery rate of one particular infectious disease. δ measures the disease transmissibility within the population. σ measures the disease transmissibility between populations

points $I_{ij}(t)$ over a time interval $[0, T]$ for the i_{th} subpopulation in the j_{th} replicate. We choose a time point s and use the data points $I_{ij}[0], I_{ij}[1], \cdots, I_{ij}[s]$ as the training data and predict the data points $I_{ij}(t)$ for $s < t \leq T$. $s = 80$ is chosen in this experiment. The MAPE of the prediction is defined as

$$E_{pre} = \frac{1}{100 * 20 * m} \sum_{j=1}^{100} \sum_{i=0}^{m-1} \sum_{s=81}^{100} \left| \frac{\hat{I}_{ij}[t] - I_{ij}[t]}{I_{ij}[t]} \right|. \quad (31)$$

The MAPE of the prediction is 17.3 %, which indicates that our model can achieve around 82.7 % accuracy in terms of the prediction. Again, we randomly select one replicate and show the prediction results in Fig. 3.

Real application

In the case study, we apply our model to the surveillance data from the 2009 H1N1 pandemic in Hong Kong. We acquired the time series data of the daily number of confirmed H1N1 cases with symptom onset from May 1, 2009 to May 23, 2010. The database includes 36 547 confirmed cases with demographic information on location, age, and sex along with the laboratory confirmation dates. The epidemic curve of confirmed H1N1 cases (see Fig. 4) reaches its peak at the end of September 2009, after which the intervention procedure comes into effect and the curve goes down. We use the data up to September 30, 2009, which comprises 27 898 cases (more than 2/3 of all cases). Hong Kong is geographically divided by 18 political areas (districts). Each district

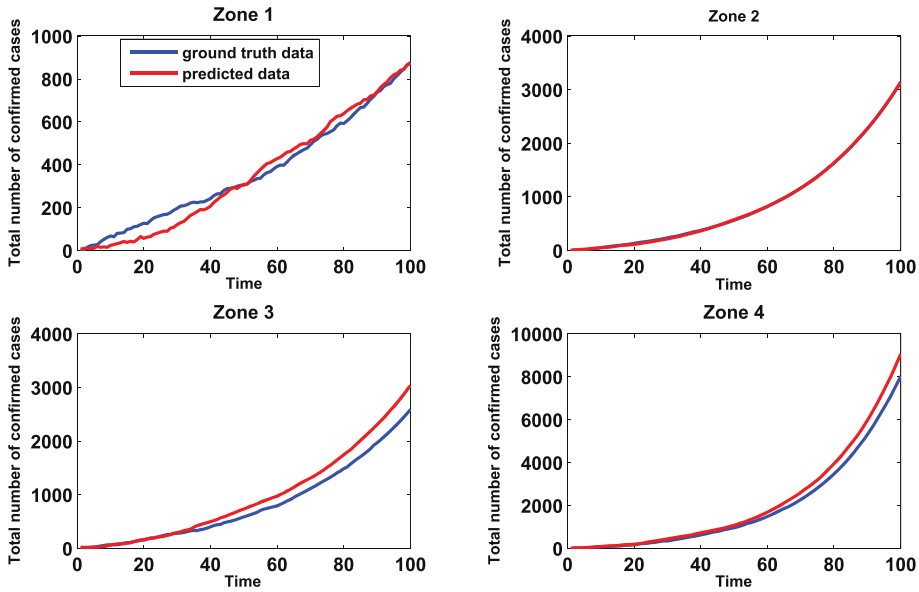

Fig. 2 The comparison of original data and estimated data for four regions. The x-axis represents the time in days. The y-axis represents the total number of confirmed cases. The The original data is generated using the ground truth values of parameters while the estimated data is generated with estimated values of parameters. The correlation between them is 0.96

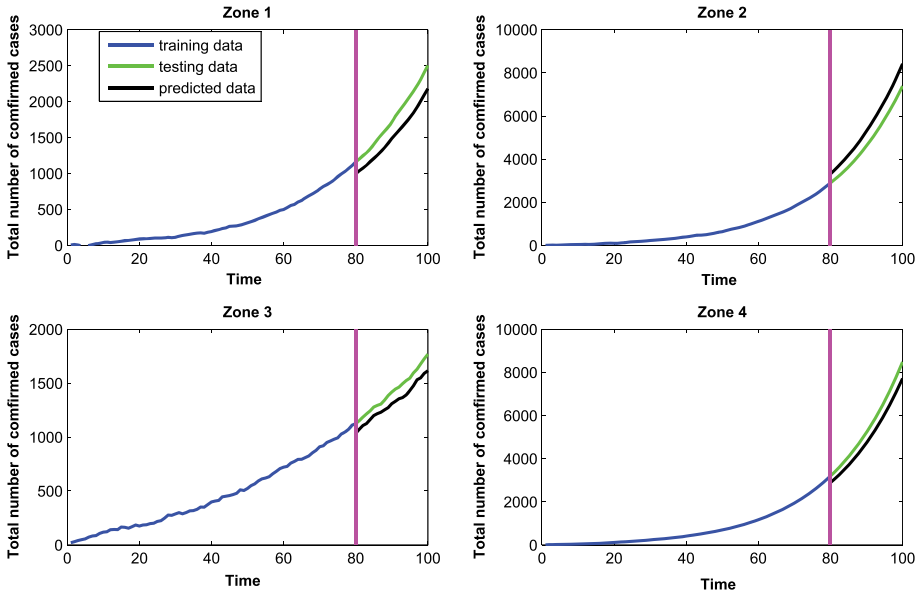

Fig. 3 The prediction performance of our method using simulation data for four regions.The x-axis represents the time in days. The y-axis represents the total number of confirmed cases. The data for the first 80 days are used for training. The data for the last 20 days are used for testing. The mean absolute percentage error in testing is 17.3 %

is considered as one sub-population in our proposed model. The time interval $\triangle t(k)(k = 0, 1, \cdots, n-1)$ of H1N1 is set as 1 day. The total number of days is 100.

Figure 5 gives the effect of the different components for the 18 political areas in Hong Kong. From Fig. 5, we can find that the effect of δ, which measures the internal disease spread within each district, varies less than the effect of σ, which measures the external disease spread between districts. In general, the speed of internal disease spread is closely connected with the population density and the external disease spread pattern is associated with the pattern of people's daily travel. It is well known that Hong Kong has the highest population density in the world, and most districts are densely populated. However, it possesses a heavy heterogeneous traffic pattern, and there is intensive traffic between districts every day. Therefore, the imported infections for each district account for a critical factor in the disease spread, whereas the internal effects only play a very small role in the progression of disease spread.

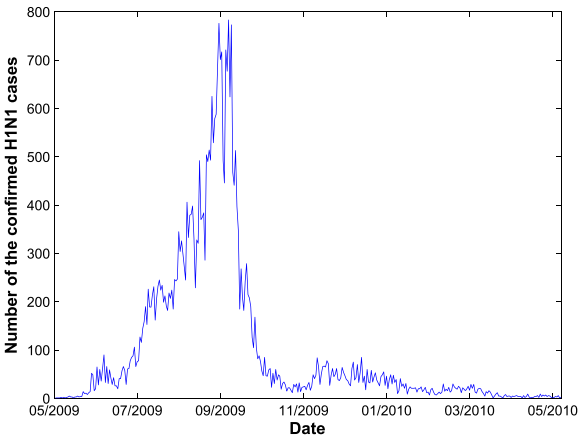

Fig. 4 Daily H1N1 epidemic curve in Hong Kong from May 1, 2009 to May 23, 2010. The epidemic curve of confirmed H1N1 cases reaches its peak at the end of September, 2010

We also use the H1N1 data to test the prediction accuracy of our model. The MAPEs for all districts are shown in Fig. 6 and Table 1. The average prediction error is 20.0 %. We notice that the prediction error for the district "TSUEN WAN" is very high because the number of daily infections in this district changes suddenly during the epidemic period. Figure 7 shows the epidemic curves of the three regions with the lowest incidence rate. We can observe that between time slot 34 and 42, there is a sudden rise for the "TSUEN WAN" district. Such a change significantly affects the parameter estimation and thereby the prediction accuracy for the district "TSUEN WAN". Although the incidence rates of the other two districts also low, their epidemic curves are relatively smooth in comparison with that of "TSUEN WAN", indicating that the prediction accuracies of these two districts are higher than that of the "TSUEN WAN" district.

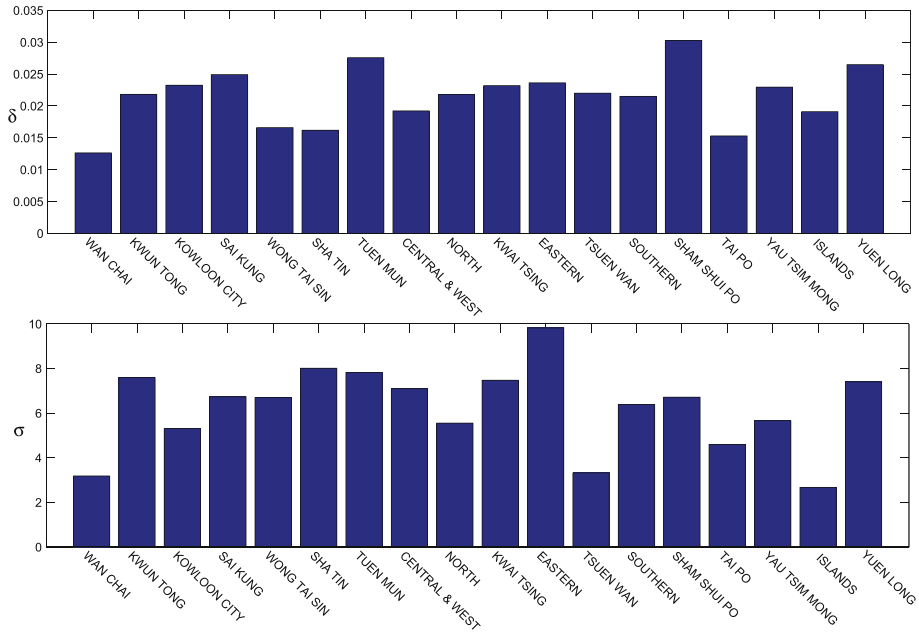

Fig. 5 The estimation of two factors in disease spread for 18 districts for 2009 H1N1 pandemic in Hong Kong. δ measures the transmissibility of disease spread within each district. σ measures the transmissibility of disease spread from the neighbors of each region. This figure shows that the imported infections for each district account for a critical factor in the disease spread while the internal effects only play a very small role in the progression of disease spread

Conclusions

Epidemic modelling offers a practical means for policy makers to evaluate the potential effects of intervention strategies. To do so, the accuracy of epidemic modelling with respect to the real-world disease transmission dynamics is essential and remains a challenging task due to the inaccessibility of many factors that affect the spread patterns of infectious diseases. In particular, heterogeneity should be taken into consideration when modelling the disease spread in non-random mixing populations. Many methods have been proposed to deal with heterogeneity in the study of epidemic dynamics, mostly using network-based epidemic models in which nodes correspond to spatial locations with reported incidences over time, and the directional links indicate the probability of disease transmission from one node to another over time. However, it is very challenging to determine the network topology. Many studies have used a geographical topology whereas others have used a mobility network inferred from the public transportation network or other sources. How to verify the inferred network topology is another challenging issue because the true epidemic network topology is unknown, and it may vary for different types of infectious diseases for the same population. Furthermore, the neighborhood effect estimation is non-trivial; it involves many parameters (a polynomial of the number of nodes) and requires a large amount of data to avoid overfitting. Such data may not always be available for the inference of network topology. Therefore, in this work, we propose an alternate approach to investigate the spatial heterogeneity from temporal-spatial surveillance data without the inference of network topology.

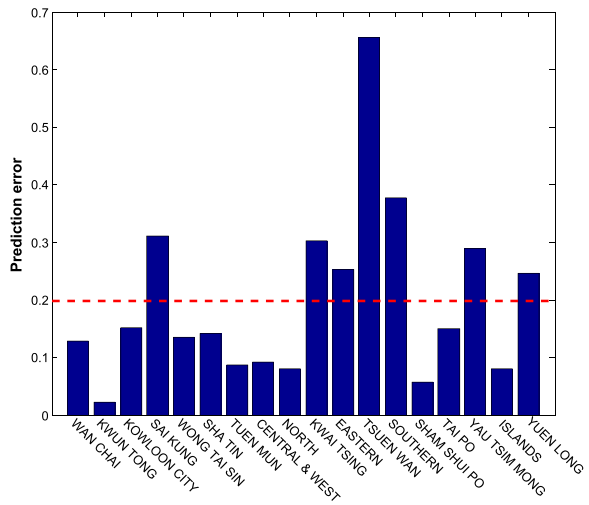

Fig. 6 The prediction errors for 18 districts using the real H1N1 data. The mean absolute percentage error is 20.0 %

Our proposed model possesses several merits over the previous works. First, it quantifies the role of the heterogeneity in the analysis of the spread dynamics

Table 1 Prediction error for real data

District	WAN CHAI	KWUN TONG	KOWLOON CITY	SAI KUNG	WONG TAI SIN	SHA TIN
X Error	0.12	0.02	0.15	0.31	0.13	0.14
District	TUEN MUN	CENTRAL &WEST	NORTH	KWAI TSING	EASTER	TSUEN WAN
Error	0.09	0.09	0.08	0.30	0.25	0.65
District	SOUTHERN	SHAM SHUI PO	TAI PO	YAU TSIM MONG	ISLANDS	YUEN LONG
Error	0.38	0.06	0.15	0.29	0.08	0.25

The average error of 18 districts is 0.20

of infectious diseases in heterogeneous populations. Second, parameter estimation can be computed very quickly. Therefore, the prediction and the corresponding intervention policies can be implemented without delay in an outbreak of infectious disease. We apply our model on both the simulated data and the real data from the 2009 H1N1 epidemic in Hong Kong and achieve acceptable prediction accuracy. Based on the study of disease diffusion, the model proposed in this work can be extended to study other propagation patterns such as the Internet and World Wide Web, through which individuals form multiple communities in which information can propagate in a manner similar to that of infectious disease. We believe that our work makes theoretical and empirical contributions in many areas.

There are some limitations in our proposed stochastic model. First, it does not consider the epidemic network topology. However, how to infer such networks is another challenging task. To the best of our knowledge, the best way to do so is to use the contact data among some infected patients to verify the results, but such data are not always available and can be difficult to collect due to many issues (e.g., privacy). This issue may be addressed by using other types of data, such as daily commute data extracted from social networks. Second, our proposed model achieves a prediction accuracy of only around 80 %. We need to further improve it to allow its full use in real applications. Third, the proposed model is only suitable for the situation in which the susceptible population (or sub-population) maintains a relatively constant size and structure in a region. However, if the number of infected people in an epidemic is large or asymptomatic infection plays a central role (e.g., the malaria epidemic in Africa), the population factor should be taken into consideration in the model. Moreover, for a highly spatially heterogeneous outbreak (e.g., the Ebola epidemic) in which cases may seem to disappear due to reduced transmission in one area while growth may continue or rise in new locales, our proposed model may have problems in capturing these opposite dynamics in different regions. Fourth, because the proposed model is based on the classic SIR model, it only works in the situation in which the number of infected people grows exponentially. We will investigate resolutions to these limitations in our future work.

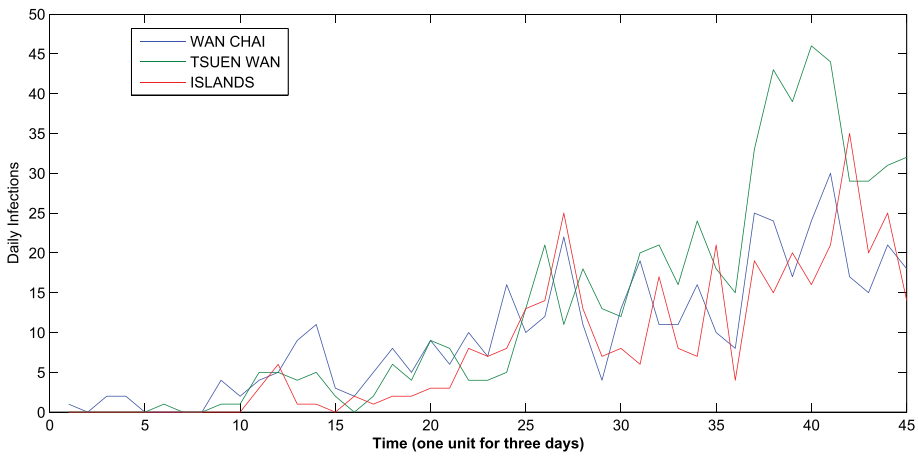

Fig. 7 The epidemic curve of three districts with the lowest incidence rates. The x-axis represents the time staring from May 1, 2009 to May 23, 2010. Between time slot 34 and 42, there is a sudden rise for the "TSUEN WAN" district. Such a change significantly affects the parameter estimation and thereby the prediction accuracy for the district "TSUEN WAN"

Additional file

Additional file 1: Multilingual abstracts in the five official working languages of the United Nations. (PDF 598 kb)

Acknowledgements
This work is supported by Hong Kong Baptist University Strategic Development Fund and Hong Kong General Research Grant HKBU12202114.

Authors' contributions
RXM, JML, and XW conceived and designed the experiments. RXM implemented the software. JML and XW analysed the data. All authors were involved in the manuscript preparation. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Statistics & Mathematics, Zhejiang Gongshang University, Hangzhou, China. [2]Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong. [3]Department of Computer Science & Institute of Theoretical and Computational Study, Hong Kong Baptist University, Kowloon Tong, Hong Kong.

Received: 21 December 2015 Accepted: 10 October 2016
Published online: 22 December 2016

References

1. World Health Organization Pandemic (H1N1). 2009. http://www.who.int/csr/don/2010_01_08/en/index.html. Accessed 1 Aug 2010.
2. Cohen J, Enserink M. As swine flu circles globe, scientists grapple with basic questions. Science. 2009;324(5927):572–3.
3. Ferguson NM, Cummings DA, Fraser C, Cajka JC, Cooley PC, Burke DS. Strategies for mitigating an influenza pandemic. Nature. 2006;442(7101):448–52.
4. Germann TC, Kadau K, Longini IM, Macken CA. Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci. 2006;103(15):5935–940.
5. Bailey NTJ, et al. The Mathematical Theory of Infectious Diseases and Its Applications. High Wycombe: Charles Griffin and Company Ltd; 1975.
6. Anderson RM, May RM. Infectious Diseases of Humans vol. 1. Oxford: Oxford University Press; 1991.
7. Heesterbeek J. Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation. vol 5. New York City: Wiley; 2000.
8. Li MY, Muldowney JS. Global stability for the SEIR model in epidemiology. Math Biosci. 1995;125(2):155–64.
9. Kuznetsov YA, Piccardi C. Bifurcation analysis of periodic SEIR and SIR epidemic models. J Math Biol. 1994;32(2):109–21.
10. Hethcote HW. Qualitative analyses of communicable disease models. Math Biosci. 1976;28(3):335–56.
11. Becker NG, Britton T. Statistical studies of infectious disease incidence. J R Stat Soc Series B Stat Methodol. 2002;61(2):287–307.
12. Ferrari MJ, Bjørnstad ON, Dobson AP. Estimation and inference of R0 of an infectious pathogen by a removal method. Math Biosci. 2005;198(1):14–26.
13. Allen L. An introduction to stochastic epidemic models. Math Epidemiol. 2008;1945:81–130.
14. Cooper BS, Pitman RJ, Edmunds WJ, Gay NJ, et al. Delaying the international spread of pandemic influenza. PLoS Med. 2006;3(6):212.
15. Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. Proc Natl Acad Sci U S A. 2004;101(42):15124–15129.
16. Hollingsworth TD, Ferguson NM, Anderson RM. Will travel restrictions control the international spread of pandemic influenza? Nat Med. 2006;12(5):497–9.
17. Keeling MJ, Woolhouse ME, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, Cornell SJ, Kappey J, Wilesmith J, Grenfell BT. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science. 2001;294(5543):813–7.
18. Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S, Burke DS. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature. 2005;437(7056):209–14.
19. Longini IM, Nizam A, Xu S, Ungchusak K, Hanshaoworakul W, Cummings DA, Halloran ME. Containing pandemic influenza at the source. Science. 2005;309(5737):1083–1087.
20. Pastor-Satorras R, Vespignani A. Epidemic dynamics and endemic states in complex networks. Phys Rev E. 2001;63(6):066117.
21. Pastor-Satorras R, Vespignani A. Epidemics and immunization in scale-free networks. 2002. arXiv preprint cond-mat/0205260.
22. Kuperman M, Abramson G. Small world effect in an epidemiological model. Phys Rev Lett. 2001;86(13):2909–912.
23. Newman ME, Jensen I, Ziff R. Percolation and epidemics in a two-dimensional small world. Phys Rev E. 2002;65(2):021904.
24. Boguná M, Pastor-Satorras R, Vespignani A. Absence of epidemic threshold in scale-free networks with degree correlations. Phys Rev Lett. 2003;90(2):028701.
25. Grassberger P. On the critical behavior of the general epidemic process and dynamical percolation. Math Biosci. 1983;63(2):157–72.
26. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature. 1998;393(6684):440–2.
27. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–12.
28. Erdős P, Rényi A. On the evolution of random graphs. Publ Math Inst Hungar Acad Sci. 1960;5:17–61.
29. Pastor-Satorras R, Vespignani A. Epidemics and immunization in scale-free networks. In: Handbook of graphs and networks: from the genome to the internet. Hoboken: Wiley Online Library; 2005. p. 111–130.
30. Rogers EM. Diffusion of Innovations. New York: Simon and Schuster; 2010.
31. Leskovec J, Adamic LA, Huberman BA. The dynamics of viral marketing. ACM T Web. 2007;1(1):5.
32. Kumar R, Novak J, Raghavan P, Tomkins A. On the bursty evolution of Blogspace. World Wide Web. 2005;8(2):159–78.
33. Leskovec J, McGlohon M, Faloutsos C, Glance NS, Hurst M. Patterns of cascading behavior in large blog graphs. In: SDM, vol. 7. SIAM; 2007. p. 551–6.
34. Salathé M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH. A high-resolution human contact network for infectious disease transmission. Proc Natl Acad Sci. 2010;107(51):22020–2025.
35. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton JF, Quaggiotto M, Van den Broeck W, Régis C, Lina B, et al. High-resolution measurements of face-to-face contact patterns in a primary school. PloS ONE. 2011;6(8):23176.
36. Dickison M, Havlin S, Stanley HE. Epidemics on interconnected networks. Phys Rev E. 2012;85(6):066109.
37. Hancock K, Veguilla V, Lu X, Zhong W, Butler EN, Sun H, Liu F, Dong L, DeVos JR, Gargiullo PM, et al. Cross-reactive antibody responses to the 2009 pandemic H1N1 influenza virus. N Engl J Med. 2009;361(20):1945–1952.
38. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, Leung GM, Ho LM, Lam TH, Thach TQ, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science. 2003;300(5627):1961–1966.
39. Mills CE, Robins JM, Lipsitch M. Transmissibility of 1918 pandemic influenza. Nature. 2004;432(7019):904–6.
40. Birrell PJ, Ketsetzis G, Gay NJ, Cooper BS, Presanis AM, Harris RJ, Charlett A, Zhang XS, White PJ, Pebody RG, et al. Bayesian modelling to unmask and predict influenza A/H1N1pdm dynamics in London. Proc Natl Acad Sci. 2011;108(45):18238–18243.
41. Towers S, Geisse KV, Zheng Y, Feng Z. Antiviral treatment for pandemic influenza: Assessing potential repercussions using a seasonally forced SIR model. J Theor Biol. 2011;289:259–68.
42. Cauchemez S, Valleron AJ, Boelle PY, Flahault A, Ferguson NM. Estimating the impact of school closure on influenza transmission from sentinel data. Nature. 2008;452(7188):750–4.
43. Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, Kent CK, Finelli L, Ferguson NM. Household transmission of 2009 pandemic influenza A (H1N1) virus in the United States. N Engl J Med. 2009;361(27):2619–627.
44. Lessler J, Reich NG, Cummings DA. Outbreak of 2009 pandemic influenza a (h1n1) at a New York city school. N Engl J Med. 2009;361(27):2628–636.

45. Klick B, Nishiura H, Ng S, Fang VJ, Leung GM, Peiris JM, Cowling BJ. Transmissibility of seasonal and pandemic influenza in a cohort of households in Hong Kong in 2009. Epidemiology (Cambridge). 2011;22(6):793.
46. Yang Y, Sugimoto JD, Halloran ME, Basta NE, Chao DL, Matrajt L, Potter G, Kenah E, Longini IM. The transmissibility and control of pandemic influenza a (H1N1) virus. Science. 2009;326(5953):729–33.
47. Jin Z, Zhang J, Song LP, Sun GQ, Kan J, Zhu H. Modelling and analysis of influenza a (H1N1) on networks. BMC Public Health. 2011;11(Suppl 1):9.
48. World Health Organization. Ebola Situation Reports. 2015. Available at: http://apps.who.int/ebola/en/current-situation/ebolasituation-report. Accessed 29 June 2015.
49. Camacho A, Kucharski A, Aki-Sawyerr Y, White MA, Flasche S, Baguelin M, Pollington T, Carney JR, Glover R, Smout E, et al. Temporal changes in Ebola transmission in sierra leone and implications for control requirements: a real-time modelling study. PLoS Curr. 2015;7:PMC4339317.
50. Chowell D, Castillo-Chavez C, Krishna S, Qiu X, Anderson KS. Modelling the effect of early detection of Ebola. Lancet Infect Dis. 2015;15(2):148–9.
51. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman J. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. J Theor Biol. 2004;229(1): 119–26.
52. Chowell G, Viboud C, Hyman JM, Simonsen L. The Western Africa Ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. PLoS Curr. 2015. doi:10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261.
53. Fisman D, Khoo E, Tuite A. Early epidemic dynamics of the West African 2014 Ebola outbreak: estimates derived with a simple two-parameter model. PLoS Curr. 2014. doi:10.1371/currents.outbreaks.89c0d3783f36958d96ebbae97348d571.
54. Gomes MF, y Piontti AP, Rossi L, Chao D, Longini I, Halloran ME, Vespignani A. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. PLoS Curr. 2014. doi:10.1371/currents.outbreaks.cd818f63d40e24aef769dda7df9e0da5.
55. Althaus CL. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. PLoS Curr. 2014. doi:10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288.
56. Webb G, Browne C, Huo X, Seydi O, Seydi M, Magal P. A model of the 2014 Ebola epidemic in West Africa with contact tracing. PLoS Curr. 2014. doi:10.1371/currents.outbreaks.846b2a31ef37018b7d1126a9c8adf22a.
57. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, Hewson R, García-Dorival I, Bore JA, Koundouno R, et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. Nature. 2015;524(7563):97–101.
58. Chowell G, Nishiura H. Transmission dynamics and control of Ebola virus disease (EVD): a review. BMC Med. 2014;12(1):196.
59. van Dijk A, Aramini J, Edge G, Moore KM. Real-time surveillance for respiratory disease outbreaks, Ontario, Canada. Emerg Infect Diseases. 2009;15(5):799.