ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Theoretical Biology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Journal of Theoretical Biology

www.elsevier.com/locate/yjtbi

A susceptible-infected model of early detection of respiratory infection outbreaks on a background of influenza

Mojdeh Mohtashemi[a,b,*], Peter Szolovits[b], James Dunyak[a], Kenneth D. Mandl[c,d]

[a] *MITRE Corporation, Cambridge, MA, USA*
[b] *MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA*
[c] *Children's Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA*
[d] *Harvard Medical School, USA*

Received 5 October 2005; received in revised form 17 January 2006; accepted 26 January 2006

Abstract

The threat of biological warfare and the emergence of new infectious agents spreading at a global scale have highlighted the need for major enhancements to the public health infrastructure. Early detection of epidemics of infectious diseases requires both real-time data and real-time interpretation of data. Despite moderate advancements in data acquisition, the state of the practice for real-time analysis of data remains inadequate. We present a nonlinear mathematical framework for modeling the transient dynamics of influenza, applied to historical data sets of patients with influenza-like illness. We estimate the vital time-varying epidemiological parameters of infections from historical data, representing *normal* epidemiological trends. We then introduce simulated outbreaks of different shapes and magnitudes into the historical data, and estimate the parameters representing the infection rates of anomalous deviations from normal trends. Finally, a dynamic threshold-based detection algorithm is devised to assess the timeliness and sensitivity of detecting the irregularities in the data, under a fixed low false-positive rate. We find that the detection algorithm can identify such designated abnormalities in the data with high sensitivity with specificity held at 97%, but more importantly, early during an outbreak. The proposed methodology can be applied to a broad range of influenza-like infectious diseases, whether naturally occurring or a result of bioterrorism, and thus can be an integral component of a real-time surveillance system.

Keywords: SI model; Transients; Early detection; Infectious disease; Outbreaks; Biosurveillance

1. Introduction

The global health, threatened by emerging infectious diseases, pandemic influenza, and biological warfare, is becoming increasingly dependent on the rapid acquisition, processing, integration and interpretation of massive amounts of data. In response to these pressing needs, new information infrastructures are needed to support active, real time surveillance. Critical for real time surveillance are two components: data collection and data analysis. Indeed today, there are several real time outbreak monitoring systems in place in major metropolitan cities (Mandl et al., 2004; Tsui et al., 2003; Lewis et al., 2002; Lober et al., 2002; Reis et al., 2003). Despite such progressive efforts, the state of the practice for detecting temporal abnormalities in surveillance data is not adequate.

For highly infectious agents, such as SARS or smallpox, a few infected individuals can propagate the infection at a rate that might initially elude the attention of public health authorities, thereby gaining time for the infected pool to grow silently but rapidly to the point after which public health measures will prove ineffective. Therefore, the challenge for any model of outbreak detection lies in the early recognition of such exponentially growing processes, when the exponential nature of the process is not easy to recognize.

Abbreviations: CH ED, Children's Hospital Emergency Department; ED, Emergency Department

[*] Corresponding author. MIT Computer Science, 77 Massachusetts Avenue, Cambridge, MA, USA.

E-mail addresses: mojdeh@mitre.org, mojdeh@mit.edu (M. Mohtashemi).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.01.031

With respect to an infectious agent, a population consists of those who are susceptible, infected or immune to the disease. Depending on the clinical and epidemiological properties of the disease there may be other categories. Here, we model the short-term dynamic interaction between different subpopulations with respect to an infectious disease using a nonlinear system of difference equations. Such a system will comprise meaningful demographic and epidemiological parameters, representing transient properties of the underlying dynamics, which can be estimated from historic epidemiological data. The resulting body of information will be the basis for defining *normality* in epidemiological trends and therefore can be used to detect anomalous deviations from historically observed events early on. We make the key assumptions that such disease processes are highly infectious, highly contagious, and manifested with non-specific flu-like symptoms in patients early in the development.

The methodology presented here represents a fusion among the developing methods of syndromic surveillance (Mandl et al., 2004; Tsui et al., 2003; Lewis et al., 2002; Lober et al., 2002; Reis et al., 2003; Greenko et al., 2003; Goldenberg et al., 2002), the notion of transients in health and disease (Mohtashemi and Levins, 2001; Mohtashemi, 2001) and well-established concepts and approaches in mathematical epidemiology (Bailey, 1967; Edelstein-Keshet, 1988; Anderson and May, 1992).

2. Methods

The data set used for this study consists of the daily number of patients presenting to the emergency department (ED) of a large urban, academic pediatric or children hospital (CH) with respiratory syndromes during the period 6/1/1992–5/31/2003. Fig. 1 illustrates the historical time series during the 5-year period 6/1/1996–5/31/2000. ED chief complaints were used to select encounters for infectious respiratory illness that are highly reflective of population patterns of influenza (Bougeois et al., in press; Brownstein et al., 2005). Chief complaint codes were chosen during the triage process, from a pre-defined on-line list of 181 choices. A previously validated subset of the constrained chief complaint set was chosen a priori for inclusion in the respiratory syndromic grouping (Beitel et al., 2004). Institutional review board approval was obtained. It is important to note that the syndromic respiratory definition closely corresponds with influenza activity as shown in Brownstein et al. (2005), which further strengthens our key assumptions about the historical data representing flu-like illnesses.

2.1. Model

Consider the following first-order nonlinear system of difference equations:

$$S_{n+1} = S_n - \beta S_n I_n,$$
$$I_{n+1} = I_n + \beta S_n I_n - \delta I_n, \tag{1}$$

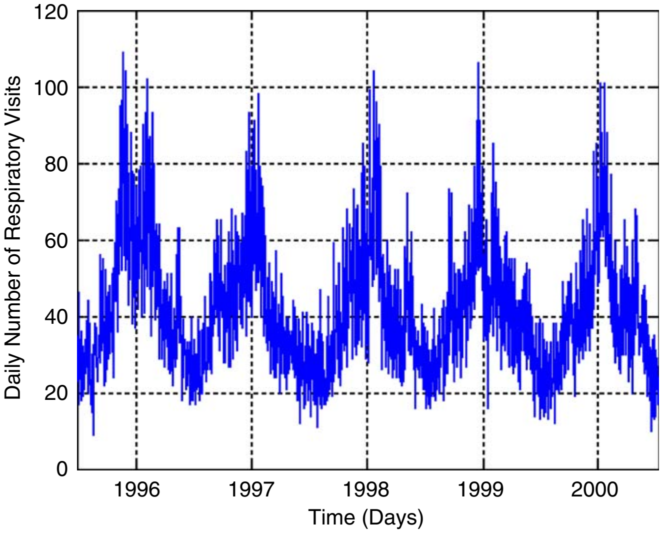

Fig. 1. Daily number of visits with respiratory infections to the CH ED during 6/1/1996–5/31/2000.

where S_n and I_n represent the respective number of susceptible and infected individuals at time n; β is the infection transmission rate; and δ is the average rate of recovery from infection so that $1/\delta$ is the mean duration of infectivity in days. Because we are interested in modeling transients in the population transmission dynamics of flu-like illnesses, implicit in our assumption is that there is not enough time for the temporary removal of the recovered population to be of significance to the dynamics. Delays introduced by temporary recovery and return to the susceptible class, although common in flu-like infections, are typically much longer than the infectivity period. Furthermore, there is no explicit representation for the latent population and the latency period in our model.

2.2. System transformation

Understanding population susceptibility is key to the control and prevention of epidemics of infectious disease. However, population susceptibility is difficult to account for. That is, the proportion or number of people susceptible at a point in time cannot be observed or measured systematically. On the other hand, the proportion or number of people infected at a point in time is an observable variable. Transforming an epidemiologically sensible system of two variables S_n and I_n, into an equation with one variable I_n, for which there are data, eliminates the unobservable variable while preserving the underlying dynamics (Mohtashemi, 2001). If we apply the elimination process to the above system of equations, then we end up with a second-order equation in the variable representing the infected population:

$$I_{n+2} = \frac{I_{n+1}^2}{I_n} - \beta I_{n+1}(I_{n+1} - (1-\delta)I_n). \tag{2}$$

The data on the number infected can then be used to estimate the unknown system parameters. Clearly, the number of children who visited the Children's Hospital Emergency Department (CH ED) on day n does not constitute the total number of infected children on day n, who live in proximity of, and seek care at the CH ED. However, if we assume an average of d days of infectivity per patient, i.e. $1/\delta = d$, then the overall number infected on day n can be approximately represented by the sum of the number of visits to the ED during the past d days, including day n. That is, for $i = d \ldots 365$ we have

$$I_i = \sum_{j=i-d+1}^{i} v_j. \tag{3}$$

Here v_j represents the number of patients presenting to the ED with respiratory infections on day j, and as observed in the CH ED data set. For $i = 1 \ldots d-1$ the summation spans the last d–i days of the previous year and the first i days of the current year. In other words, we use a d-day sliding window to approximate the number of infected children for the last day in the window. For example, if the average infectivity period $1/\delta = d = 7$, then the number of infected children on day 7 can be approximated by the sum of the number of children who presented to the CH ED with influenza during the past 7 days including day 7. These numbers, approximately, reflect the population of infected children who live in the catchments area of CH and come to the ED when ill. Such a method of approximation, using a moving window of ED visitors, seeks to compensate for variation in health-seeking behavior of a population due to various socio-demographic factors including holidays and day of the week. It is important to note that the number of infected children, estimated according to the ED data and using Eq. (3), is representative of all infected individuals who seek treatment at the ED. Implicit in this statement is the assumption that the underlying disease process causes fairly serious symptoms, and thus generates active health-seeking behavior.

If in Eq. (2), we take I_{n+2} to represent the number infected today then the term $(1-\delta)I_n$ represents the infected number from two days ago who remained infected yesterday, and $I_{n+1}-(1-\delta)I_n$ represents the number who presented with respiratory infection at the ED yesterday. The term $\beta I_{n+1}(I_{n+1}-(1-\delta)I_n)$ then represents the number who cannot contribute to new infections for today since they are already infected. The number infected today, I_{n+2}, is therefore a nonlinear function of the total numbers infected during the past 2 days, the ratio in the right-hand-side of Eq. (2), offset by an adjustment term.

2.3. Estimating the infection rate

Epidemiologically vital parameters of infectious diseases are seldom constant. It is only over short time periods that the constancy assumption may be valid. Furthermore, despite the temporal variation within each year, it is quite likely that the epidemiological properties of shorter time periods remain relatively stable across different years. Defining a suitable time window during which system properties remain relatively stable depends on many factors including observed periodicities in the data and demographic and epidemiologic properties of the disease under study. Examples include seasonal, bimonthly, monthly, biweekly, and weekly time windows.

To arrive at average estimates for β representative of change in the historical data, the daily number of visits for every year in the data set was converted to the daily number infected according to Eq. (3), and fed into Eq. (2) in a sliding manner resulting in 365 such equations for every year in the historical data. We assumed that the mean duration of infectivity is 7 days, i.e. $1/\delta = 7$, a clinically sensible assumption for influenza (Steinhoff, 2001; Wearing et al., 2005). Next, we chose a time window of length l for which a system of l–2 such equations was formed, and β was estimated using least-squares regression. We then iteratively calculated β for each time-window throughout each year by sliding the time window to arrive at 365 different values of β corresponding to 365 sliding time windows, one for each day of the year. In this paper, the results for $l = 7$ are reported. This is a sensible choice for the window since it compensates for the day of the week variation in the data set. Fig. 2 illustrates the daily mean infection rates estimated using the training set (see Section 2.4).

Because β follows the dynamics of change in the infected population within each day's short past history, it can be used to detect deviation from normal more sensitively than the actual daily values if the average behavior of β under normal circumstances can be modeled. But more importantly, because β reacts to unexpected change almost

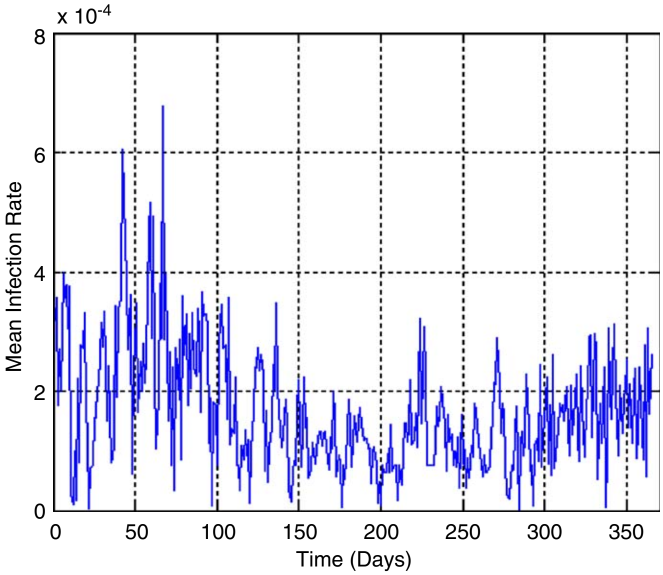

Fig. 2. Mean daily infection rates estimated from historical training data during the 7-year period, 6/1/1992–5/31/1999.

instantly, this property can be used to detect abnormalities in the data quite early on during an outbreak. This is an important property of the proposed methodology because the ability to detect early is a principal factor differentiating the effectiveness of outbreak detection systems.

2.4. Model validation

The first 7 years in the historical data were used as training years and the remaining 4 years as test years. The training years were used to estimate average values of β corresponding to 365 days in a year. We then introduced simulated outbreaks into every year in the test set, estimated the corresponding values of β under outbreak condition for each test year. The mean and standard deviation of β for each day of the year across 7 years in the training set were used to detect deviation from normal.

2.5. Simulated outbreaks

For every test year in the historical data set we introduced 7-day long, 30-day apart, simulated outbreaks into the actual ED daily visit rates. Each outbreak was defined by a controlled feature set enabling testing of the system over a range of outbreak types. Every month in the data set contained a 7-day outbreak. Such semi-synthetic data sets consisted of 48 outbreaks throughout the 4 years in the test set. We also generated outbreaks randomly throughout the test data set and compared the results with those of deterministically simulated outbreaks. In this paper, we report the results from the deterministic simulations because first, the results from the two experimental simulations were comparable, and second, our results can be compared to those from a previous study by Reis et al. (2003), where they generated outbreaks every 15 days throughout the historical data. The inclusion of one outbreak in every month in the test set was to assess performance of the methodology against different values of background incidence of respiratory reports, including seasonal, and to avoid the biasing effects of estimating β during an outbreak on estimates of β during non-outbreak days since the time windows can potentially overlap.

Because we were interested in testing the detection capability of the model under outbreaks that start slowly and follow an exponential growth, for the base outbreak simulations we chose a concave up exponential function that would not only generate relatively small numbers during the first 4 days (<20) but the sum of the numbers infected during the first 5 days would not exceed 75. (For a mathematical treatment of the choice of epidemic trajectories we refer the reader to Appendix A.) These numbers were chosen relative to the daily mean and standard deviation across the training and test years so that the results across different models of outbreak as well as from a previous study by Reis et al. (2003) can be compared, and to have a metric for early detection. In this paper, an outbreak is considered to be detected *late* if it is detected during the last 2 days of the outbreak.

The concave up exponential function that we used for simulating outbreaks is of the form $round(4.1e^{0.385k})$, $k = 1 \ldots 7$. This function generates the numbers 6, 9, 13, 19, 28, 41, and 61 that are, respectively added to a 7-day long window of daily visits that are designated as outbreak days. The generated numbers are smaller than 20 during the first 4 days, slightly higher on the 5th day, and the sum of simulated numbers during the first 5 days is 75. Although in real epidemics of highly contagious infectious disease the cumulative numbers infected during the first week may far exceed such slow growth, our goal is to be able to detect early, when in fact the actual numbers are easily hidden in the baseline.

Although a concave up exponential function produces the most realistic shapes for outbreaks of contagious infectious disease, we also implemented two other functions simulating two different shapes of outbreak in order to allow for a comprehensive comparative study. A concave down exponential function of the form $round(23(1-e^{-0.4k}))$, $k = 1 \ldots 7$, was used to generate the numbers 8, 13, 16, 18, 20, 21, and 22 respectively during a 7-day long outbreak window, where the numbers during the first 4 days are less than 20 and the sum of the numbers over the first 5 days is 75. Although the total number infected during the 7-day period is different under each model, we only consider the differential detection power between the models during the first 5 days because the numbers during the last 2 days under the exponential models are too high to be missed even without any analysis. The concave down exponential model of outbreak is well suited for early detection because the rate of change in the earlier part of the dynamics is faster than that of the later, and thus it can be used as the best case scenario to test the early detection capability of our framework. Finally, we added a constant of size 15 to all 7 days of each simulated outbreak to represent a uniform model of outbreak.

2.6. Detection threshold

For each day n in the training years, the mean infection rate $\bar{\beta}_n$, and standard deviation σ_n were estimated. In conjunction with a numerical threshold, these values were used to determine the strength of change in the newly estimated parameter β_n, for each year in the test set infused with simulated outbreaks. On day n the detection algorithm raises an alarm and counts that day as an outbreak day if $(\beta_n \geqslant \bar{\beta}_n + 2\sigma_n)$ & $(\beta_n > T)$, where $T = 2 \times 10^{-4}$ is a numerical lower bound and $n = 1 \ldots 365$. The detection threshold was set so that the detection system would generate an average of 3.3% false positives per year in the training set devoid of simulated outbreaks. This is an empirical adjustment to make the specificity the same across all experiments, so that their sensitivity results can be compared fairly. Furthermore, this is likely to be a reasonable assumption and manageable

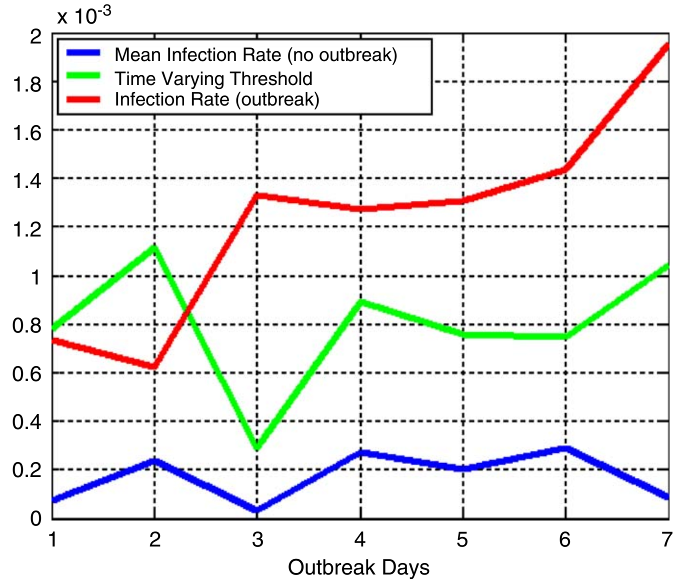

Fig. 3. Dynamics of the mean infection rate $\bar{\beta}_n$, the time varying detection threshold $\bar{\beta}_n + 2\sigma_n$, and the infection rate β_n, during a randomly generated seven-day long outbreak in the test set. The outbreak is detected on the third day when $(\beta_3 > \bar{\beta}_3 + 2\sigma_3)$ and $(\beta_3 > T)$.

performance rate for surveillance systems, and it has been adopted in the literature (Reis et al., 2003). Fig. 3 illustrates the dynamics of the mean infection rate $\bar{\beta}_n$, the time-varying detection threshold $\bar{\beta}_n + 2\sigma_n$, and the estimated infection rate in the test set β_n, during a randomly generated 7-day long outbreak in the test set. Note that the outbreak cannot be detected on either the 1st or the 2nd day since $\beta_i < \bar{\beta}_i + 2\sigma_i$ for $i = 1, 2$. The outbreak is finally detected on the third day where we have $(\beta_3 > \bar{\beta}_3 + 2\sigma_3)$ and $(\beta_3 > T)$.

3. Results

3.1. Sensitivity and specificity

We used a sliding 7-day detection window for estimating the parameter β for each day of the year in the test set infused with simulated outbreaks of different shape and size. We assumed the mean duration of infectivity is 7 days (see system of Eqs. (1)), a clinically reasonable assumption for flu-like respiratory infections. For every day the corresponding β was derived and compared to the detection threshold for that day. There were a total of 48 outbreaks generated throughout the 4 years in the test set. Table 1 reports the overall sensitivity under different models of outbreak, where overall sensitivity is defined as the number of detected outbreaks divided by the total number of simulated outbreaks. Of the 48 (100%) simulated outbreaks we detected 42 (87.5%) under the concave up model of outbreak, 33 (69%) outbreaks under the concave down model, and 30 (62.5%) under the uniform model of outbreak. The 95% confidence intervals were estimated using standard Gaussian assumptions.

Table 1
Overall sensitivity, with 95% confidence interval (95% CI), of the detection system under different models of outbreak reported as the number of detected outbreaks over the total number of outbreaks, $k = 1 \ldots 7$

Type of outbreak	Overall sensitivity	95% CI
Concave up: *round*(4.1 exp(0.385*k*))	0.875 (42/48)	0.866, 0.884
Concave down: *round*(23(1−exp(−0.4*k*)))	0.688 (33/48)	0.66, 0.72
Uniform: 15	0.625 (30/48)	0.58, 0.67

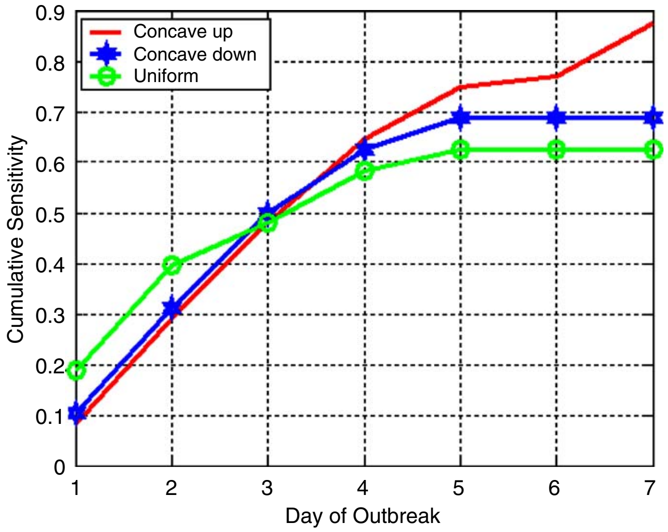

Fig. 4. Timeliness of detection based on cumulative sensitivity under different models of outbreak.

Fig. 4 demonstrates the cumulative daily sensitivity of the detection system under different models of outbreaks, where daily sensitivity is defined as the number of detected outbreaks on each day that is designated as an outbreak day, divided by the total number of outbreaks. Specificity was fixed at 97% for all experiments so that the sensitivity results can be compared (see Section 2.6).

Finally, the overall sensitivity of the model can also be measured under different values for number of false positives. This is an important performance measure because it provides insight into the tradeoff between sensitivity and specificity. Fig. 5 is generated under the SI model applied to the test data infused with exponentially concave up outbreaks. The mean false alarm rate varies from 0 to 0.06 (about 22 false alarms per year).

3.2. Timeliness of detection

Together Fig. 4 and Table 1 identify an interesting property of the detection system. As long as there is a change in the dynamics of the observed data, the detection algorithm continues to detect that change. This is perhaps more evident from the distribution of sensitivity under the least realistic model of outbreaks of infectious disease, the uniform. Under the uniform model of outbreak, although

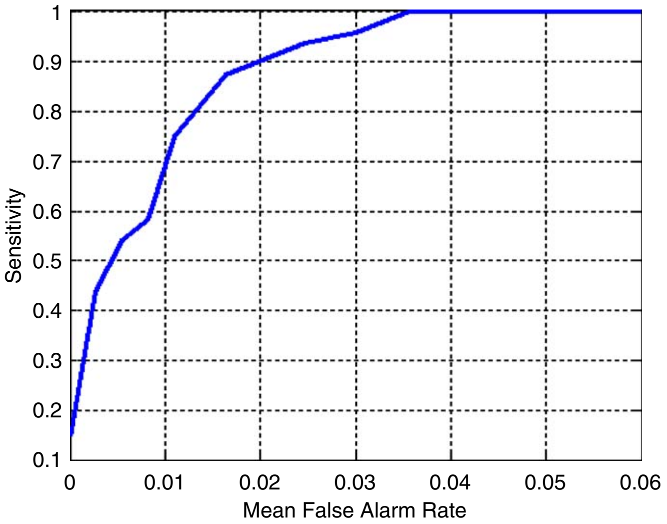

Fig. 5. Tradeoff between sensitivity (detection probability) and mean false alarm rate under the exponentially concave up model of outbreak.

the overall sensitivity is lower than the exponential models, its daily sensitivity is higher on the first 3 days. This is a sensible result and supports the premise that β reacts instantly to unexpected change, and that it helps to detect anomalies by tracking the change within a short time window. On the first day, there are 15 more visitors causing β to adjust immediately. Thus, about 18% of all simulated outbreaks (over 30% of all detected outbreaks) are spotted on the 1st day. On the other hand, because β follows the change in the dynamics of data it will report no change after the 4th day since the same number is being added from day 1 on. A similar property is observed under the concave down exponential model of outbreak. Because the rate of change is faster during the earlier days than that of later days of the dynamics of the simulated function, the detection system spots about 63% of all simulated outbreaks (about 95% of all detected outbreaks) by day 5, after which the system identifies very little change.

Under both exponential models of outbreak over 63% of all simulated outbreaks are detected during the first 4 days of outbreak. Prior to the 4th day of outbreak, the system detects more outbreaks under the exponential concave down model of outbreak compared to the concave up model of outbreak. This is because the change in the dynamics of the concave down model during the first few days of the outbreak is faster than those under the concave up model of outbreak, and therefore the exponential nature of the outbreak is recognized earlier. After the 3rd day of the outbreak, at which point the daily sensitivities of the two models coincide, the system detects more outbreaks under the most realistic model of outbreak, the concave up exponential. This is because after a relatively slow growth during the first 2 days of outbreak the detection system reacts immediately when the early signs of the exponential nature of the outbreak are recognized. It is important to note that under the concave up exponential model of outbreak, despite the lower than mean simulated number of visits during the first 2 days, nearly 30% of all simulated and detected outbreaks are spotted by the 3rd day.

Although the cumulative sensitivity of the system under the concave up exponential model is relatively lower than the other two models of outbreak during the first 3 days, the system catches over 85% of all detected outbreaks (about 75% of all simulated outbreaks) during the first 5 days so that not only does it eventually detect most of the simulated outbreaks (88%), but it detects the majority of them before day 6.

3.3. *Model performance and time of year*

Measures of model performance vary depending on the time of year and are not uniformly distributed throughout each year in the data. We therefore classified the measure of daily sensitivity into three categories: *early*, represented as the sum of daily sensitivities during the first 3 days of outbreak, *intermediate*, represented as the sum of daily sensitivities during the 4th and 5th days of outbreak, and *late*, represented as the sum of daily sensitivities during the last 2 days of outbreak. We divided each year into four seasons and for every season in each year in the test set the category of sensitivity was determined under the basic exponential concave up model of outbreak. Fig. 6 embodies the seasonal sensitivities under the three different sensitivity categories, averaged over 4 years in the test set.

The winter season, December through February, has the lowest measure of early sensitivity and highest measure of late sensitivity. This is because when the respiratory related number of visits are well on the rise or relatively high, historically during November through early January, and very high, historically in mid-late January and February, adding the same numbers to the background of positively steep slopes and already high numbers makes them less

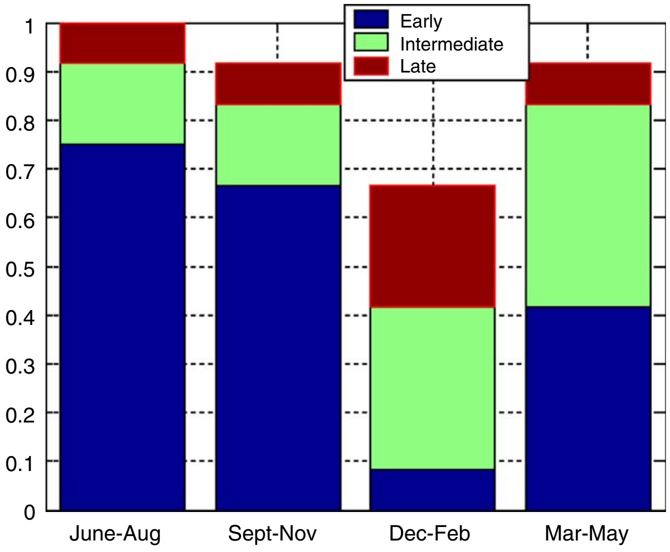

Fig. 6. Seasonal sensitivity: timeliness of detection during four seasons averaged over 4 years in the test set. Outbreaks were generated according to the exponentially concave up model.

visible to the detection algorithm than otherwise. During the summer season, June–August, when respiratory-related visits to the ED are historically the lowest, adding even relatively small numbers to the daily numbers make them more visible to the detection algorithm, thereby attaining the highest measure of early sensitivity. Hence, a real time surveillance system must consider such biases in the effect of seasonality on the power of detection in the data analysis and development of detection algorithms.

3.4. *Model performance and recovery period*

Thus far, we have examined the sensitivity of our detection algorithm under the assumption that $1/\delta = 7$, which is clinically a sensible assumption for flu-like respiratory infections (Steinhoff, 2001; Wearing et al., 2005). Here, we examine the effect of varying values of recovery period on the overall sensitivity of the detection system. Fig. 7 illustrates the sensitivity profile of the detection system as the recovery period is increased. If we rearrange Eq. (2) we get

$$I_{n+2} = \frac{I_{n+1}^2}{I_n} - \beta I_{n+1}(I_{n+1} - I_n) - \delta\beta I_n I_{n+1}. \tag{4}$$

As the recovery period takes on larger values, δ becomes smaller causing I_{n+2} in Eq. (4) to be overestimated and thus more discernable to the detection algorithm, which leads to higher detection sensitivity. On the other hand, as the recovery period decreases, δ takes on larger values causing I_{n+2} to be underestimated and less visible to the detection algorithm, leading to lower detection sensitivity.

3.5. *Seasonally adjusted detection thresholds*

The results of Figs. 4 and 6 can be improved upon by generating multiple detection thresholds depending on the time of year under analysis. The biasing effect of seasonality on lowering the detection power of the algorithm during the late fall and early winter months can be reduced if the detection algorithm is trained under multiple detection thresholds. We extended the algorithm for generating detection thresholds from the previous section so that on day n the detection algorithm raises a flag and counts that day as an outbreak day if

$$\begin{array}{ll} (\beta_n \geqslant \bar{\beta}_n + 2\sigma_n + T_1) & \text{if } n \in \{\text{November}, \ldots, \text{February}\}, \\ (\beta_n \geqslant \bar{\beta}_n + 2\sigma_n)\&(\beta_n > T_2) & \text{otherwise,} \end{array}$$

where $T_1 = 0.5 \times 10^{-4}$ and $T_2 = 2 \times 10^{-4}$ are numerical bounds and $n = 1 \ldots 365$.

Fig. 8 illustrates the results of application of the seasonally adjusted detection algorithm under the three models of outbreak. Clearly, if an outbreak occurs during the switch from the first period to the second, this approach may introduce glitches in the system. However, any detection algorithm, if adopted in a real time surveillance system, would have to be adjusted to accommodate practical matters that may not have been foreseen when developing the methodology. For instance, the problem with discrete periods can be addressed by running the algorithm under both thresholds during all or parts of the respective last and first months of the two periods. Alternatively, there might be a continuous approximation to these models that will do the smoothing automatically.

Although the qualitative properties of the results of Fig. 8 are quite similar to those of Fig. 4 with a single detection threshold, both the overall and daily sensitivities of the system have improved considerably. Under the concave up exponential model of outbreak, despite the lower than mean simulated number of visits during the first 2 days, nearly 40% of all simulated outbreaks are spotted by the 3rd day. In contrast to Fig. 4, under the concave up

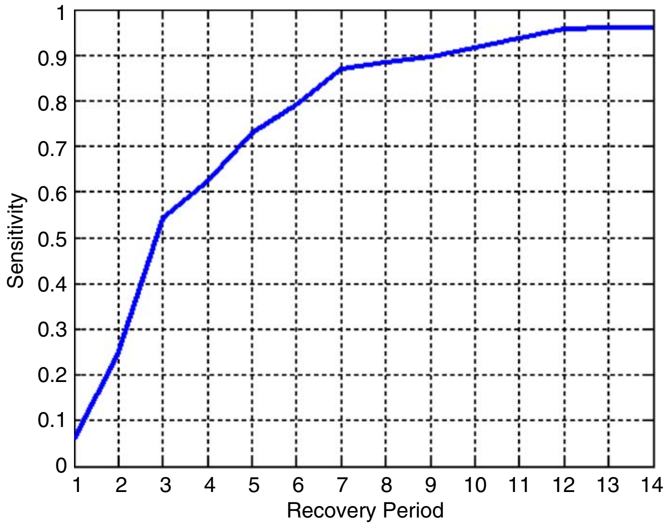

Fig. 7. Tradeoff between sensitivity (detection probability) and recovery period under the exponentially concave up model of outbreak.

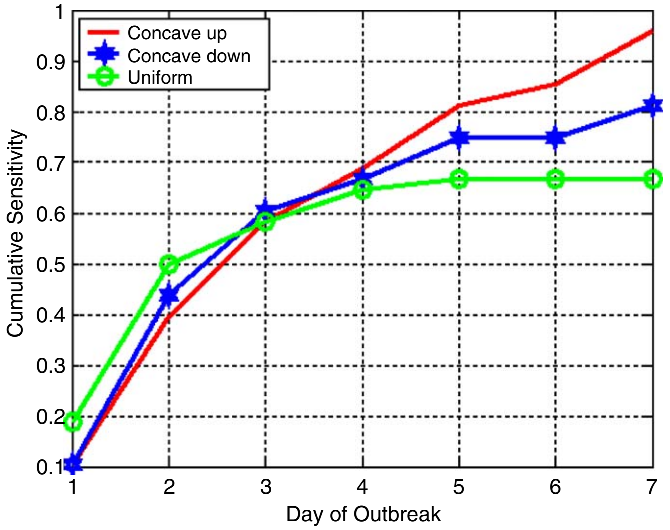

Fig. 8. Timeliness of detection based on cumulative sensitivity under different models of outbreak. The detection algorithm is modified with additional detection threshold to compensate for the biasing effect of seasonality on the detection power when the numbers are on the rise and high during mid fall and winter.

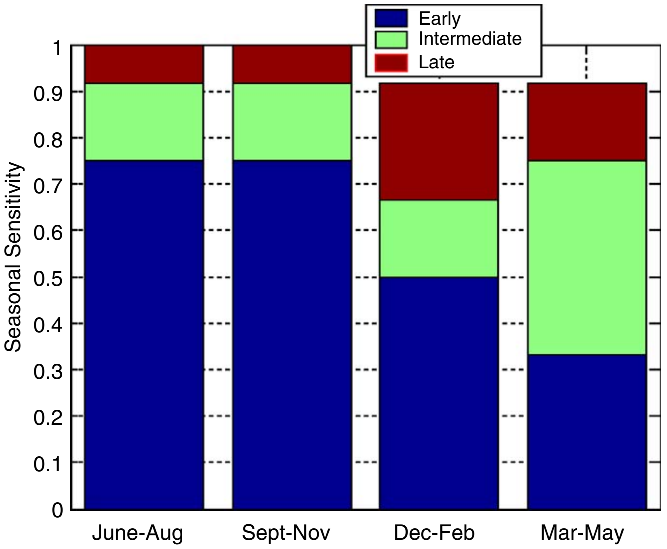

Fig. 9. Seasonal sensitivity: timeliness of detection during 4 seasons averaged over four years in the test set. Sensitivity was measured using seasonally adjusted thresholds. Outbreaks were generated according to the exponentially concave up model.

model of outbreak the sensitivity on day 1 matches that of the concave down. Note that under the concave up model of outbreak about 30% of all detected outbreaks are detected on day 2 alone. Although the cumulative sensitivity of the system under the concave up exponential model of outbreak is lower than the uniform model of outbreak during the first 2 days, the system catches over 80% of all simulated outbreaks during the first 5 days so that not only does it eventually detect most of the simulated outbreaks (96%), but it detects the majority of them before day 6. Fig. 9 embodies the seasonal sensitivities under the three different sensitivity categories defined previously, averaged over 4 years in the test set.

3.6. Further examination of detection timeliness

In this paper, we explored the idea that there is structure in the short term properties of the underlying dynamics in the CH ED data that can be captured using nonlinear models of infectious disease for timely detection. Here, we further explore this idea by comparing the sensitivity and timeliness of the SI model with those of a simple model, both applied to the CH ED test data infused with exponentially concave up outbreaks.

The model chosen for comparison is a variation of the "historical limits" (HL) method adopted and discussed in Stroup et al. (1989), Centers for Disease Control and Prevention (1989), Centers for Disease Control and Prevention (1991), and Hutwagner et al. (2003). In the HL method, the total number of reported cases in the current 4-week period is compared against the mean plus twice the standard deviation of 15 four-week intervals, spanning the 13 four-week intervals from the preceding 5 years, the current and the immediately preceding 4-week intervals.

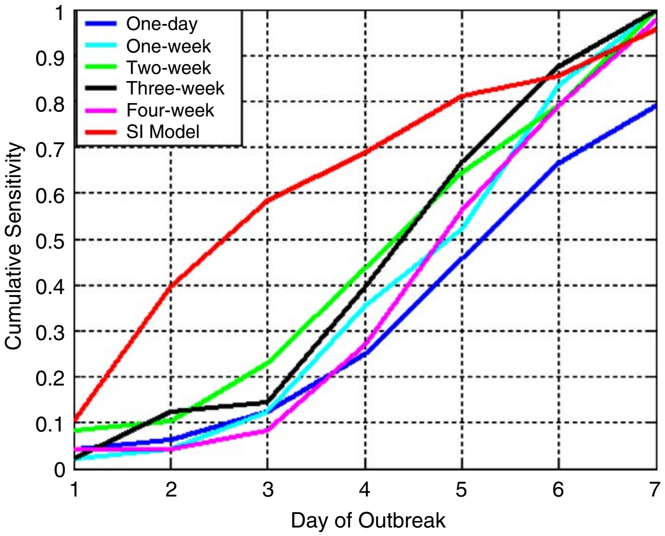

Fig. 10. Comparison of sensitivity and timeliness of the SI detection model with five variants of the HL method.

Fig. 10 illustrates the cumulative sensitivity and timeliness of five variations of the HL method and the SI model. The five variations are based on comparison of the daily counts with the mean plus a scalar times the standard deviation of a fixed interval, spanning the interval immediately preceding the current day and the same period from the preceding 5 years. The simulation results in Fig. 10 are those of the 1-day, 1-week, 2-week, 3-week and 4-week interval. The scalar in each case was selected to achieve 1.6% false-positive rate under exponentially concave up simulated outbreaks in the test set. The range of the scalar in the simulations in Fig. 10 is 2–3.8.

It is interesting to note that except for the 1-day method, all other variations of the HL method detect over 95% of the simulated outbreaks. However, none of the five variations of the HL method is suited for early detection. The best performances are those of the 2-week model where nearly 40% of the simulated outbreaks are detected during the first 4 days, and the 3-week model where nearly 67% of outbreaks are detected during the first 5 days. In contract, the performance of the SI model during the first 5 days indicates that it is more suited for early detection. Moreover, the HL model under the 4-week interval exhibits poor performance compared to the 2- and 3-week intervals. In fact, as the period of analysis is further increased past the 3-week interval, the cumulative sensitivity decreases (results not shown). This is perhaps supporting evidence for the conjecture that there is relative structure and stability in the short term dynamics of the respiratory data that should be exploited carefully for early detection, and that otherwise cannot be identified in longer term dynamics.

Overall, methods such as the HL variations examined here are reasonable models for detection when the numbers surpass a historical threshold. However, they are not designed for early detection when the numbers of excess cases are relatively low.

4. Discussion

We presented a nonlinear framework for modeling short-term transmission dynamics of influenza and early detection of anomalous events, applied to real historical data. We showed that when simulated outbreaks are introduced into the data, the detection algorithm, based on meaningful epidemiological parameters of population dynamics of infectious disease transmission and time-varying detection thresholds, is capable of detecting the irregularities in the data with high sensitivity, specificity and in a timely manner.

In the study data set, most patients come from a well-defined geographic region (the catchments area of a hospital) and are classified syndromically—based on data before confirmed diagnoses are made. Thus, inherent in the data are two properties: spatial specificity and infection non-specificity. While daily visits to the CH ED constitute 60% of all pediatric ED visits in the city, some patients seek care at other institutions. Thus the number of children infected on any day, obtained according to Eq. (3) and solely based on the CH ED statistics, is only a fraction of the total number of infected children for that day. This number is then representative of a subpopulation of infected children who have lived in the vicinity of CH and have made visits to its ED when ill. Such spatial specificity, however, is imperative for detecting abnormal events rapidly and decisively because unusual events often occur in a spatially localized manner, and are seldom distributed uniformly throughout space (Kleinman et al., 2004; Kulldorff et al., 2005).

At the same time, while the CH data comprise many different respiratory diseases with multi-factorial etiology and perhaps different clinical properties, the data encompass the same set of contagious infectious diseases that have been naturally recurring every year during the past 11 years in the data set. These infections share the property that their early clinical manifestations are similar to those of influenza. Furthermore, certain clinical and epidemiological properties of these infections appear to be similar (e.g. infectivity period). Thus, such non-specificity in disease etiology in the historical data, although appearing as limiting, can be effectively used towards detection of new disease processes of which there are no records in the historical data, but which cause similar clinical flu-like symptoms early on in patients.

A critical topic worthy of note is that of susceptibility. If the population susceptible to a disease could be systematically identified then public health policies would be more concentrated on prevention, where the health of the public is improved, instead of control, where the health of the public is maintained at a substantial cost. Despite its vitality to the problem of infectious diseases the dynamics of susceptibility is poorly understood. Reformulating an epidemiologically sensible system of two variables, i.e. susceptible and infected, into one observable variable for which we have data, i.e. infected, proves to be doubly beneficial.

On the one hand, it enables us to bypass the existing uncertainties embedded in the unobservable variable while maintaining the underlying dynamics, thereby estimating the unknown parameters that govern the system dynamics. By exploiting this property we were able to devise a detection system capable of detecting relatively small irregularities in the historic data infused with simulated outbreaks early on during the outbreak.

On the other hand, once the unknown parameters are estimated, together with the observable variable (infected), the hidden variable (susceptible) can be estimated. In the absence of birth and population influx in the system of Eqs. (1), the susceptible population, back processed and estimated, may be underrepresented. It is indeed plausible to identify the susceptible population more precisely by making the model more realistic. Such realism can be achieved by incorporating other epidemiologic and demographic parameters including age, disease-specific incubation and latency periods, the rate of loss of immunity, birth and death rates, and variables such as the latent population. The potential impacts of the outcome of such reverse processing of the data, with the goal of identifying the susceptible population, include redirecting public health policies towards preventive measures, thereby reducing morbidity and mortality due to outbreaks of infectious disease and reducing their substantial cost to society.

Some of the limitations of the proposed method need to be addressed. One such limitation is that our susceptible-infected model is not suited for analysis of non-contagious disease processes, such as anthrax, since it is designed to capture the transmission dynamics embedded in the interaction between the infected and susceptible populations. Although, it is quite conceivable that effective real time surveillance may have to rely on a combination of different detection algorithms that perform differently under various outbreak conditions and data sets, the relative power of different surveillance techniques cannot be systematically addressed until a methodical evaluation scheme is adopted and applied to various detection methods and data sets (Kulldorff et al., 2004; Dunyak et al., 2005).

Another potential limitation in our technique is in the dependence of the detection algorithm on estimating the infected population of the CH ED catchments area based on the observed number of daily ED patients. We approximated the daily number of people infected in the area by Eq. (3) using a time window as long as the average infectivity period, which is an epidemiologically reasonable assumption. Clearly, such approximations were further validated by the application of Eqs. (2) and (3) to the CH ED data set as demonstrated by our results. However, the goodness of such approximations can potentially impact the accuracy of early detection results.

The framework presented here is robust with high detection sensitivity and specificity which can thus be used to define the basis for naturally occurring epidemiological events pertaining to flu-like illnesses and to detect deviation from historically observed trends. The model can be potentially applied to a broad range of contagious infectious diseases manifested with non-specific flu-like symptoms, naturally occurring such as West Nile and SARS, or maliciously instigated such as smallpox.

Acknowledgments

MM was supported in part by NIH contract N01-LM-3-3515 and National Library of Medicine grant R01 LM007677-01. PS was supported in part by NIH grant 1-U54-LM 08748 and NIH contract N01-LM-3-3515. KM was supported by National Library of Medicine grant R01 LM007677-01 and Alfred P. Sloan Foundation grant 2002-12-1.

Appendix A

A.1. Epidemic trajectories

The choice of epidemic trajectory for outbreak simulations is a critical issue in the development of detection algorithms. Consider the system of difference equations (1) introduced in Section 2. Without loss of generality, suppose that S_n and I_n represent the respective proportion of the population who are susceptible and infected. At the early stages of an outbreak, a very small fraction of the population has been infected so that for a small time period n from the start of the outbreak we have $S_n \approx S_0$, where S_0 represents the proportion of the susceptible population at the start of the outbreak. Then we have

$$I_{n+1} \approx (1 + \beta S_0 - \delta) I_n. \tag{5}$$

Thus the increase in the infected population at time $t = n + m$ is

$$I_{n+m} \approx (1 + \beta S_0 - \delta)^m I_n. \tag{6}$$

We can rewrite Eq. (6) as

$$\begin{aligned} \frac{I_{n+m}}{I_n} &\approx (1 + \beta S_0 - \delta)^m \\ &\approx \left(1 + \frac{\beta S_0 - \delta}{m}\right)^m. \end{aligned} \tag{7}$$

This is because at the start of an outbreak $S_0 \approx 1$, and thus $|\beta S_0 - \delta| \ll 1$. Finally, application of the result $e^x = \lim_{n\to\infty}(1 + x/n)^n$ to approximation (7) suggests an exponential model of the epidemic trajectory.

References

Anderson, R.M., May, R.M., 1992. Infectious Diseases of Humans. Oxford Science Publications, Oxford.

Bailey, N.T.J., 1967. The Mathematical Approach to Biology and Medicine. Wiley, New York.

Beitel, A.J., Olson, K.L., Reis, B.Y., Mandl, K.D., 2004. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. Pediatr. Emerg. Care 20 (6), 355–360.

Bougeois, F., Olson, K.L., Brownstein, J., McAdams, A., Mandl, K.D, in press. Validation of syndromic surveillance for respiratory infections. Ann. Emerg. Med., in press.

Brownstein, J.S., Kleinman, K.P., Mandl, K.D., 2005. Identifying pediatric age groups for influenza vaccination using a regional surveillance system. Am. J. Epidemiol. 162 (7), 686–693.

Centers for Disease Control and Prevention, 1989. Proposed changes in format for presentation of notifiable disease report data. MMWR Morb. Mortal Wkly Rep. 38, 805–809.

Centers for Disease Control and Prevention, 1991. Update: graphic method for presentation of notifiable disease data—United States 1990. MMWR Morb. Mortal Wkly Rep. 40, 124–125.

Dunyak, J, Mohtashemi, M., Kulldorff, M., 2005. Benchmarking Temporal Surveillance Techniques. In: Proceedings of the Syndromic Surveillance Conference, September 2005.

Edelstein-Keshet, L., 1988. Mathematical Models in Biology. The Random House, Inc., New York.

Goldenberg, A., Shmueli, G., Caruana, R.A., 2002. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proc. Natl. Acad. Sci. USA 99, 5237–5240.

Greenko, J., Mostashari, F., Fine, A., Layton, M., 2003. Clinical evaluation of the Emergency Medical Services (EMS) ambulance dispatch-based syndromic surveillance system, New York City. J. Urban Health 80 (2 Suppl. 1), i50–i56.

Hutwagner, L., Thompson, W., Seeman, G.M., Treadwell, T., 2003. The bioterrorism preparedness and response: early aberration reporting system (EARS). J. Urban Health 80 (2 Suppl. 1), i89–i96.

Kleinman, K., Lazarus, R., Platt, R., 2004. A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. Am. J. Epidemiol. 159 (3), 217–224.

Kulldorff, M., Zhang, Z., Hartman, J., Heffernan, R., Huang, L., Mostashari, F., 2004. Benchmarking data and power calculations for evaluating disease outbreak detection methods. MMWR 53 (Suppl.), 144–151.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., Mostashari, F., 2005. A space-time permutation scan statistic for disease outbreak detection. PLoS Medicine 2 (3), 1–9.

Lewis, M.D., Pavlin, J.A., Mansfield, J.L., O'Brian, S., Boomsma, L.G., et al., 2002. Disease outbreak detection system using syndromic data in the greater Washington, DC area. Am. J. Prevent. Med. 23 (3), 180–186.

Lober, W.B., Karras, B.T., Wagner, M.M., Overhage, J.M., Davidson, A.J., et al., 2002. Roundtable on bioterrorism detection: information system-based surveillance. J. Am. Med. Inf. Assoc. 9 (2), 105–115.

Mandl, K.D., Overhage, J.M., Wagner, M.M., Lober, W.B., Sebastiani, P., et al., 2004. Implementing syndromic surveillance: a practical guide informed by the early experience. J. Am. Med. Inf. Assoc. 11 (2), 141–150 PrePrint published Nov 121, 2003 as.

Mohtashemi, M., 2001. A new method for estimating the parameters of disease dynamics. Presented at the International Conference on Mathematical and Theoretical Biology.

Mohtashemi, M., Levins, R., 2001. Transient dynamics and early diagnostics in infectious disease. J. Math. Biol. 43, 446–470.

Reis, B.Y., Pagano, M., Mandl, K.D., 2003. Using temporal context to improve biosurveillance. Proc. Natl. Acad. Sci. USA 100, 1961–1965.

Steinhoff, M.C., 2001. Epidemiology and prevention of influenza. In: Nelson, K.E., Williams, C.M., Graham, N.M.H. (Eds.), Infectious Disease Epidemiology: Theory and Practice. Aspen Publishers, Gaithersburg, MD, pp. 477–494.

Stroup, D.F., Williamson, G.D., Herndon, J.L., Karon, J., 1989. Detection and aberrations in the occurrence of notifiable disease surveillance data. Stat. Med. 8, 323–329.

Tsui, F.-C., Espino, J.U., Dato, V.M., Gesteland, P.H., Hutman, J., et al., 2003. Technical description of RODS: a real-time public health surveillance system. J. Am. Med. Inf. Assoc. 10, 399–408.

Wearing, H.J., Rohani, P., Keeling, M.J., 2005. Appropriate models for the management of infectious diseases. PLoS Med. 2 (7), e174.