# Detection of Outbreaks in Syndromic Surveillance Data Using Monotonic Regression

Jared Burdin[†], Mojdeh Mohtashemi[†§],
Martin Kulldorff[‡], James Dunyak[†]
[†]The MITRE Corporation
[§]MIT Computer Science and AI Lab
[‡]Harvard Medical School and Harvard Pilgrim Health Care

## Abstract

Due to nonstationarity and substantial variability in outbreak profiles, early detection of disease outbreaks is challenging. In this paper we propose a method to detect outbreaks in syndromic surveillance data using a generalized likelihood ratio test in which both the null and alternative hypotheses are normally distributed. The data is daily counts of interactions between patients and the National Bioterrorism Syndromic Surveillance Demonstration Program System in the Boston area. Using Poisson regression, we estimate the daily means and variances of the data as well as day of the week variations. The estimated means serve as the means under the null hypothesis. To determine the means under the alternative hypothesis we use a generalized form of the Pool-Adjacent-Violators algorithm on five-day windows of data. For each window a test statistic is computed and an outbreak is indicated if it exceeds a threshold.

**Keywords:** Syndromic surveillance, early outbreak detection, generalized likelihood ratio test, pool-adjacent-violators algorithm

## 1. Introduction

The timely detection of an infectious disease outbreak is among the most important problems facing the medical and homeland defense communities today. Effective solutions to this problem will improve the rapidity of medical response to both natural and manmade contagion, and will reduce the morbidity, mortality, and cost of an outbreak.

Efforts in this area are ongoing, with researchers exploring a wide range of methods and data sources. Recent work has become even more intense in light of the threat of bioterrorism. Goldenberg, Shmueli, Caruana, and Fienberg (2002) construct a threshold number of pharmaceutical transactions for a day, and detect an outbreak if the actual sales exceed that number. Data from 26 military treatment facilities and 99 infectious disease clinics in the Washington DC area is used by Lewis, et al (2002) to construct estimates and confidence intervals for the day's counts in several syndrome categories. Outbreaks are detected on days on which the daily count in a syndrome category exceeds its 95% confidence interval. Reis, Pagano, and Mandl (2003) study using time windows of data to improve outbreak detection sensitivity and specificity. Mohtashemi, Szolovits, Dunyak, and Mandl (2006) employ a susceptible-infected model to detect outbreak days characterized by a greater than expected infection transmission rate.

In this paper we propose a simple Generalized Likelihood Ratio Test (GLRT) to detect outbreaks in time series data. Like other techniques, the calculated test statistic is based on several days of data. It shares the benefits of other techniques in that it does not require any information on individual patients, such as their home addresses, and is generic to conventional biosurveillance data (e.g. emergency room visits), nonconventional data (e.g. over-the-counter drug sales), or some combination of the two. The GLRT differs from previous work in early detection of outbreaks in that it generates a test statistic based on the shape of the trend in the excess over the expected daily count of events (e.g. medical encounters) over some time window of data. Because the time series data should provide some indicator of the level of illness in the general population, we term the excess number of events as *excess morbidity*.

We expect outbreaks to exhibit a monotonically nondecreasing trend in the mean of the excess morbidity. We estimate this trend for time windows of data with the output of the Pool-Adjacent-Violators Algorithm (PAVA) (Härdle 1990, p. 218). The sum of the estimated mean of the excess morbidity and the estimated daily mean number of events serves as the mean of the alternative hypothesis. The mean under the null hypothesis of the GLRT is provided by the estimated daily means alone. Here the estimated daily means are provided by a Poisson regression over the entire data set.

In this paper we apply the GLRT to five years of daily counts of interactions between individuals diagnosed with an upper respiratory infection and the National Bioterrorism Syndromic Surveillance Demonstration Program system in the vicinity of Boston, Massachusetts. The performance of the GLRT is ascertained by inserting stochastic outbreaks into the data and measuring the probability of detection (i.e., sensitivity) and the probability of false alarm (i.e., 1 - specificity). The early detection capability, or *timeliness*, of the detector is also measured.

## 2. Materials and Methods

We apply the GLRT to real data in order to measure its performance. Variations in the data are accounted for using a Poisson regression. Artificial outbreaks were inserted into the data so that probability of detection and probability of false alarm could be calculated.

### 2.1 Data

The data were provided by the National Bioterrorism Syndromic Surveillance Demonstration Program (NDP) and involve ambulatory care encounters of patients using a large medical practice in eastern Massachusetts and having health insurance through a major insurer in the region (Lazarus, Kleinman, Dashevski, DeMaria, and Platt 2001; Lazarus et. al 2002; Platt et. al 2003; Yih et al. 2004). Specifically, the dataset consists of counts of new episodes of illness by date of medical encounter and by syndrome during the five-year period of January 1, 2000-December 31, 2004. "New episodes" of illnesses were those not preceded by an encounter for the same syndrome in the previous 42 days. "Encounters" or "events" included office visits, urgent care visits, and telephone calls to primary care providers. Syndromes considered were upper gastrointestinal (GI) illness, lower GI, respiratory, influenza-like illness (ILI), and neurological and were defined in terms of sets of diagnostic codes (CDC 2003). A single event could be included in the daily count of more than one syndrome if the patient had diverse symptoms (e.g., vomiting (upper GI) and diarrhea (lower GI)) or if one of his/her diagnostic codes mapped to more than one syndrome (e.g., influenza with pneumonia, which is in both ILI and respiratory syndromes). Here we focus on the upper respiratory infection syndrome.

Analysis and modeling of the data is complicated by significant nonstationarity evidenced by day of the week and seasonal variations which exist in the data set. Figure 1 illustrates the variation in the mean counts for each day of the week. The day of the week variation is likely the result of each daily count's reliance on doctor visits as a source of events; one plausible explanation for the significantly lower counts on the weekends is that most doctors' offices are closed on the weekend, and appointments due to medical conditions arising over the weekend are scheduled for the following Monday. Not surprisingly, holidays also impact the daily count. As such, we consider holidays separately than other days (a holiday falling on a Monday is considered to be a holiday, and not a Monday). Lewis et al (2002) also considers daily variation, weekly trends, and holiday effects in their analysis. However, unlike Lewis et al (2002), we do not consider after-holiday effects here. In addition, weekend days are ignored in this analysis because of the significant difference between the mean number of events on weekdays and weekends and because four weekend days are missing from the data set.

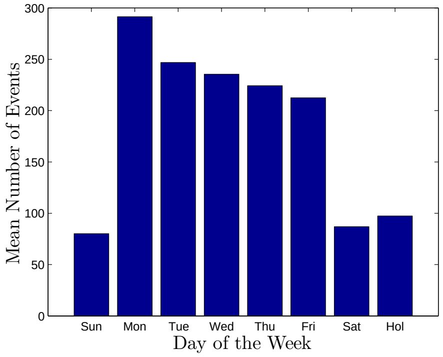

**Figure 1. Day of the Week and Holiday Variation in the NDP Upper Respiratory Infection Data**

Figure 2 shows the seasonal variation in the number of interactions between patients diagnosed with upper respiratory infections and the NDP system for weekdays in the data set. Autumn and winter have the greatest mean number of events, and spring and summer have the least, as is expected.

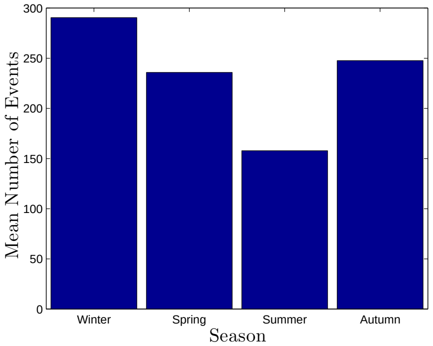

**Figure 2. Seasonal Variation in the Upper Respiratory Infection NDP Data**

### 2.2 Over-Dispersed Poisson Regression

To handle the seasonal and weekday variations, we exploit the Poisson regression to obtain $\lambda_t$ , the expected number of patients for each day, *t*, in the data set. A Poisson approximation is a natural choice to fit the data, given that the data is a counting process. Furthermore, the data manifests Poisson properties, in that the variance grows with the mean. However, the variance is greater than the mean, particularly at larger means. Thus, we fit the data to an over dispersed Poisson. The model is of the form

$$\begin{aligned}\ln\lambda_t = \ln\mu_t + \beta_1 + \beta_2 I_{Tue} + \beta_3 I_{Wed} + \beta_4 I_{Thu} \\ + \beta_5 I_{Fri} + \beta_6 I_{Hol},\end{aligned} \tag{1}$$

where $\beta_1, \beta_2, \ldots, \beta_6$ are the regression coefficients, $I_{Tue}, I_{Wed}, \ldots, I_{Hol}$ are indicator functions that indicate whether the day under consideration is a Tuesday, Wednesday, Thursday, Friday, or holiday (Monday is indicated when all the indicator functions equal zero), and $\mu_t$ is the predictor variable for day *t*. The predictor variable is calculated as the arithmetic mean number of events over a ten-day period that ends five days before day *t*.

## 2.3 Generalized Likelihood Ratio Test

Once $\lambda_t$ is calculated, $x_t$, the excess morbidity on day *t*, may be determined. The excess morbidity is expressed as

$$x_t = k_t - \lambda_t, \tag{2}$$

where $k_t$ is the daily count of new episodes of upper respiratory infection on day *t*. Here we detect outbreaks having a monotonically nondecreasing trend in mean excess morbidity. The trend in the mean excess morbidity is determined via the Pool-Adjacent-Violators Algorithm (PAVA) (Härdle 1990, p. 218), which provides a set of minimum mean squared error estimates for the data under the monotonicity constraint. In other words, the PAVA finds the set of estimates $\hat{x}_t$ such that

$$\varepsilon = \frac{1}{l}\sum_{t=\tau-(l-1)}^{\tau}\left(x_t - \hat{x}_t\right)^2 \tag{3}$$

is minimized, where *l* is the window length, and $\tau$ is the most recent day in the window (i.e., today), under the restriction that $\hat{x}_{\tau-(l-1)} \le \hat{x}_{\tau-(l-2)} \le \ldots \le \hat{x}_\tau$. If $x_t$ is normally distributed, the estimates have the additional property of being the maximum likelihood estimates for the data under the monotonicity constraint. The estimates, in addition to the corresponding $\lambda_t$, serve as means under the alternative hypothesis of the GLRT.

The GLRT is utilized to determine whether a trend in excess morbidity is due to an outbreak process or is simply the result of ordinary day-to-day variations in the data. Here, we invoke the central limit theorem in order to utilize a Gaussian approximation for the distribution of the data under the null and alternative hypotheses. Under the null hypothesis the data is distributed with mean of $\lambda_t$ and variance $\sigma_w^2$, where $\sigma_w^2$ is a function of $\lambda_t$ and the dispersion parameter $\psi$ estimated in the Poisson regression, given by

$$\sigma_w^2 = \frac{1}{l}\sum_{t=\tau-(l-1)}^{\tau}\psi\lambda_t. \tag{4}$$

The dispersion parameter is estimated by dividing the sum of weighted residuals squared (Pearson's chi-square) by the degrees of freedom. Under the alternate hypothesis the data is distributed with mean of $\left(\lambda_t + \hat{x}_t\right)$ and variance $\sigma_w^2$. In this paper, we use a window length of 5 days. Thus, the test statistic G is given by

$$\mathrm{G} = \frac{\prod_{t=\tau-4}^{\tau}\frac{1}{\sqrt{2\pi\sigma_w^2}}e^{-\frac{1}{2}\left(\frac{k_t-\lambda_t-\hat{x}_t}{\sigma_w}\right)^2}}{\prod_{t=\tau-4}^{\tau}\frac{1}{\sqrt{2\pi\sigma_w^2}}e^{-\frac{1}{2}\left(\frac{k_t-\lambda_t}{\sigma_w}\right)^2}} \tag{5}$$

and the log likelihood ratio test is given by

$$\mathrm{G}_{\ln} = \sum_{t=\tau-4}^{\tau}\frac{2\hat{x}_t\left(k_t-\lambda_t\right)-\hat{x}_t^2}{2\sigma_w^2}. \tag{6}$$

An outbreak is detected when the test statistic $\mathrm{G}_{\ln}$ is above a threshold value. Because of the variation in the data, a different threshold is calculated for each time window of the data by simulating data under the null hypothesis 4000 times, and calculating and ordering the resulting $\mathrm{G}_{\ln}$ values. The sorted $\mathrm{G}_{\ln}$ values serve as thresholds, with smaller $\mathrm{G}_{\ln}$ values corresponding to smaller *p*-values.

It should be noted that the emphasis of the test is on the trend in the mean of the excess morbidity, not its magnitude. In other words, any monotonically nondecreasing trend in the mean excess morbidity has the potential to trigger a detection, regardless of whether the excess morbidity values are greater or less than zero.

## 2.4 Artificial Outbreaks

Unfortunately, it is not possible to derive performance results from the original NDP data because the locations of actual outbreaks in the data set are not known. Therefore, the data set was augmented with artificial, stochastic outbreaks at known locations. The artificial outbreaks were randomly created from the outcomes of Poisson random variables. The means of the random variables had trends based on two families of shapes, uniform and exponential growth. Figure 3 and Figure 4 show the trends in the mean for several artificial three and five-day exponential outbreaks, respectively (the exact values of the means are shown in brackets in the figures). Figure 5 shows the trend in the mean for three artificial three-day uniform outbreaks. One day and five day uniform outbreaks were also inserted into the data; the daily means for these outbreaks were the same as those for the three-day uniform outbreaks, only the lengths of the outbreaks were different. The outbreaks were inserted by adding the outcome of the random variables to the daily counts on the desired days. To ensure that the performance measure is not influenced by the location in time of the artificial outbreaks, outbreaks were added to every five-day window of data.

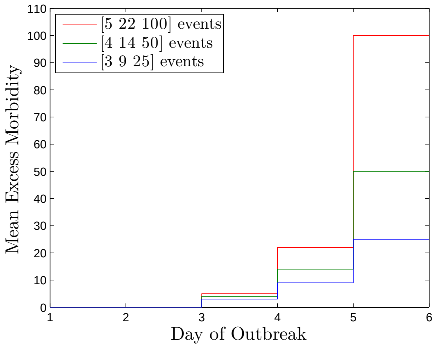

**Figure 3. Trend in the Mean of Poisson Random Variables for Artificial Three-Day Exponential Outbreaks**

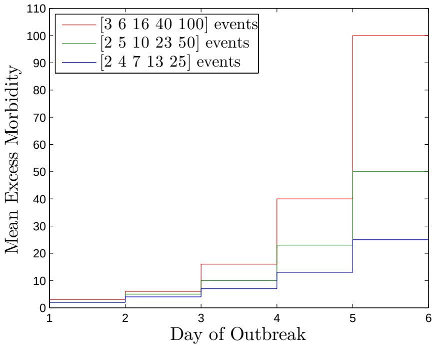

**Figure 4. Trend in the Mean of Poisson Random Variables for Artificial Five-Day Exponential Outbreaks**

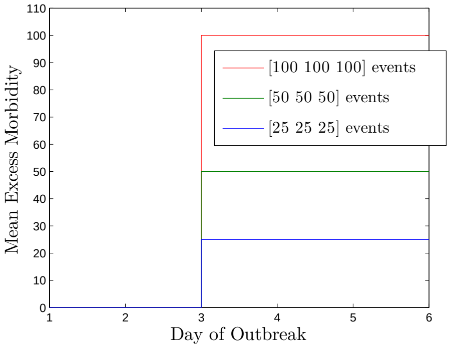

**Figure 5. Trend in Mean of the Poisson Random Variable for Artificial Three-Day Uniform Outbreaks**

## 3. Results

Receiver Operating Characteristic (ROC) curves were constructed for the detector based on the test (6) in order to determine its performance. The ROC curves for three and five-day exponential outbreaks are shown in Figure 6 and Figure 7, respectively. The ROC curves appear similar because of the means chosen for the random variables. In both cases, for each of the three levels of outbreak, the outbreaks end with the same mean number of events. Had the trends been chosen differently, different results would have been produced. However, it should be noted that the detector performs nearly as well when presented with three-day exponential outbreaks as when it is presented with five-day exponential outbreaks that end on the same mean number of events.

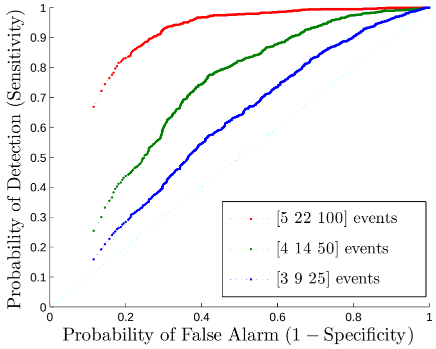

**Figure 6. ROC Curve for Three-Day Exponential Outbreaks**

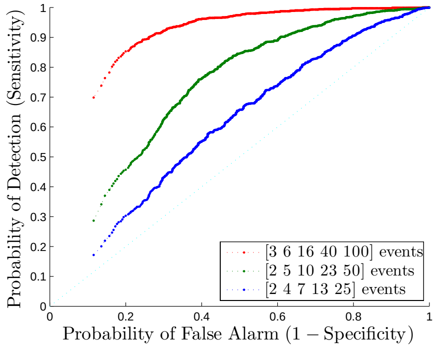

**Figure 7. ROC Curves for Five-Day Exponential Outbreaks**

The ROC curves for one, three, and five day uniform outbreaks are shown in Figure 8, Figure 9, and Figure 10, respectively. In the case of uniform outbreaks, as the figures show, detection probability is heavily influenced by outbreak length. Detection probability improves when the outbreak length increases from one day to three days in all three cases examined, particularly those for which the mean number of outbreak events per day is 100. While present, the improvement in detection probability is less pronounced between three and five-day outbreak lengths.

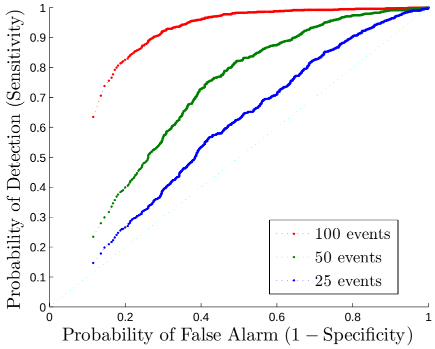

**Figure 8. ROC Curves for One-Day Uniform Outbreaks**

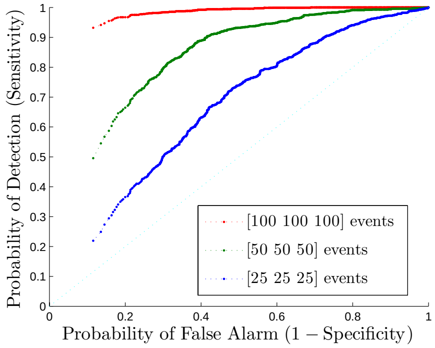

**Figure 9. ROC Curves for Three-Day Uniform Outbreaks**

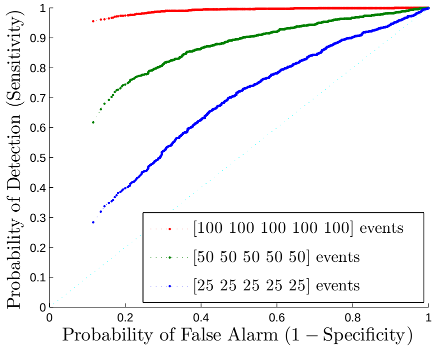

**Figure 10. ROC Curves for Five-Day Uniform Outbreaks**

In addition to measuring the probability of detection for various probabilities of false alarm, the timeliness of the detector was also investigated. We define timeliness as the ability of the detector to detect an outbreak early in the outbreak process (Reis et al. 2003). Practically, we measure timeliness by sliding the detection window across an inserted stochastic outbreak and recording each day of the outbreak a detection is signaled. That is, we begin with a detection window filled with four days of unaugmented data and one day (i.e. the last or most recent day) filled with the first day of the outbreak; the window is then moved one day at a time until the entire window is filled with outbreak data. Each window position, or outbreak day, the PAVA is run on the data in the window and the likelihood ratio test is applied. If the outcome of the test exceeds the threshold for the five days in the window for a given false alarm probability, an outbreak is indicated.

Figure 11 shows the detection rates for a false alarm probability of 0.1639 for each day of an outbreak for exponential and uniform outbreaks. In this case, the means of each day of the stochastic outbreaks were [3 6 16 40 100] in the case of exponential outbreaks, and [50 50 50 50 50] in the case of uniform outbreaks. As Figure 11 shows, the detector, when presented with an exponential outbreak having this shape, exhibits low timeliness. The timeliness of the detector in the face of uniform outbreaks is much better than in the face of exponential outbreaks. The probability of detecting the uniform outbreak is nearly twice as large as the probability of detecting the exponential outbreak on day one of the outbreaks, and more than twice as large in days 2 through 4 of the outbreaks. On the last day of the outbreaks, conversely, the probability of detection of the exponential outbreak is greater than that of the uniform outbreak.

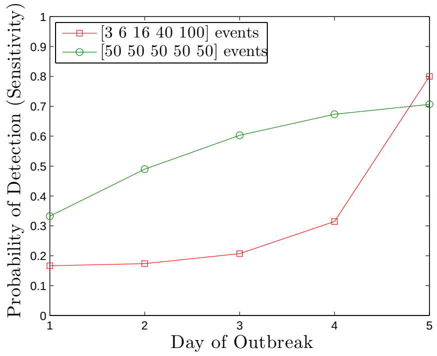

**Figure 11. Detector Timeliness**

When comparing the detection probabilities in the face of different outbreak shapes, however, it is important to note the power in the outbreak signals. Overall, the power in the mean of both outbreak shapes is approximately equal, assuming each outbreak day is independent, with the power in the mean of the exponential outbreaks being less than that of the uniform outbreaks. However, the power in the first four days of the mean of the uniform outbreak signal is approximately five times greater than that of the exponential outbreak. This undoubtedly accounts for the improved timeliness of the detector in the face of the uniform outbreaks. The improved probability of detection of the exponential outbreaks on the last day of the outbreaks is due to the fact that the signal power in day five of the mean of the exponential outbreaks is four times that of the uniform outbreaks.

## 4. Conclusion

Early detection of infectious disease outbreaks is an important and challenging problem. In this paper we present a solution based on signal processing techniques. We propose a Generalized Likelihood Ratio Test (GLRT) to detect outbreaks in time series data. Unlike prior outbreak detection techniques, the GLRT generates a test statistic based on the shape of the trend in the excess morbidity over some time window of data. In addition, it does not require any information on individual patients, a benefit shared by other techniques.

The detector could provide a real time outbreak alarm when presented with time series data from one or more data sources. The detector, matched to monotonically nondecreasing trends in mean excess morbidity, was evaluated for performance in the face of exponentially and uniformly shaped outbreaks. For a given signal strength and false alarm rate, the detector is better able to detect exponential outbreaks. On the other hand, the detector manifests increased early detection performance, or timeliness, in the face of uniform outbreaks.

## References

CDC (2003), "Syndrome definitions for diseases associated with critical bioterrorism-associated agents," unpublished manuscript, available at http://www.bt.cdc.gov/surveillance/syndromedef/index.asp.

Goldenberg, A., Shmueli, G., Caruana, R. A., and Fienberg, S. E. (2002), "Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales," *Proceedings of the National Academy of Sciences*, 99, 5237-5240.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, p. 218.

Lazarus R, Kleinman K, Dashevsky I, DeMaria A, and Platt R. (2001), "Using automated medical records for rapid identification of illness syndromes: the example of lower respiratory infection," *BioMed Central Public Health*, 1, 1–9.

Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria Jr. A, and Platt R. (2002), "Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events," *Emerging Infectious Disease*, 8, 753–760.

Lewis, M. D., Pavlin, J. A., Mansfield, J. L., O'Brien, S., Boomsma, L. G., Elbert, Y., and Kelley, P. W. (2002), "Disease Outbreak Detection System Using Syndromic Data in the Greater Washington DC Area," *American Journal of Preventive Medicine*, 23, 180-186.

Mohtashemi, M., Szolovits, P., Dunyak, J., and Mandl, K. D. (2006), "A susceptible-infected model of early detection of respiratory infection outbreaks on a background of influenza," *Journal of Theoretical Biology*, 241, 954–963.

Platt R, Bocchino C, Caldwell B, Harmon R, Kleinman K, Lazarus R, Nelson AF, Nordin JD, and Ritzwoller DP (2003), "Syndromic surveillance using minimum transfer of identifiable data: the example of the National Bioterrorism Syndromic Surveillance Demonstration Program," *Journal of Urban Health* 80 (Suppl. 1), i25–i31.

Reis, B. Y., Pagano, M., and Mandl, K. D. (2003), "Using temporal context to improve biosurveillance," *Proceedings of the National Academy of Sciences*, 100, 1961-1965.

Yih WK, Caldwell B, Harmon R, Kleinman K, Lazarus R, Nelson A, Nordin J, Rehm B, Richter B, Ritzwoller D, Sherwood E, and Platt R. (2003) "The National Bioterrorism Syndromic Surveillance Demonstration Program. In: Syndromic Surveillance: Reports from a National Conference," *Morbidity and Mortality Weekly Report* 2004, 53 (suppl), 43-46.