A Family of Temporal Surveillance Techniques for Detecting Increasing Outbreaks

James Dunyak[1], Ph.D., Mojdeh Mohtashemi[1,2], Ph.D.,

Katherine Yih[3], Ph.D., Martin Kulldorff[3], Ph.D.

[1]The MITRE Corporation; [2]MIT AI and CS Lab; [3]Harvard Medical School and Harvard Pilgrim Health Care

Abstract

Emergence of new infectious agents spreading at a global scale has added new urgency to the development of real-time disease surveillance systems. At the same time, biological terrorism is a growing concern. The successful interdiction against SARS demonstrated the confrontation of these urgent crises through rapid and accurate detection of unusual epidemiological trends. Many surveillance algorithms have been proposed in the literature, but comparing these approaches is complicated. Benchmarking of temporal surveillance techniques is a critical step in the development of an effective syndromic surveillance system. Unfortunately, holding "bake-offs" to blindly compare approaches is a difficult and often fruitless enterprise, in part due to the parameters left to the final user for tuning. In this paper, we demonstrate how common analytical development and analysis may be coupled with realistic data sets to provide insight and robustness when selecting a surveillance technique. Four detection approaches, all part of a family of detectors, are considered: G-surveillance based on the space-time scan statistic, a uniformly most powerful test for exponentially increasing outbreaks, a monotonic regression approach, and a non-negative regression approach. The first of the four is a standard technique, while the latter three are new techniques for syndromic surveillance. All are compared using time series on patient visits coded as influenza-like illness at Harvard Pilgrim Health Care in the Boston area. The tradeoff in detection capability and robustness demonstrates the benefit of a monotonic regression approach for growing outbreaks. If an exponentially growing outbreak is the target, then the uniformly most powerful test delivers nearly optimal performance.

Introduction

Successful disease surveillance and early detection are dependent on both real-time collection and timely interpretation of syndromic data. While most surveillance systems are capable of monitoring and capturing appropriate data, the detection of early outbreaks remains a considerable challenge. When outbreaks are substantially localized both temporally and spatially, then detection is enhanced. However, due to our mobile society, many (if not most) outbreaks will occur over larger geographic scales, reducing the power of spatially oriented tests. Here we focus on temporal-only testing strategies applied to a metropolitan area.

This paper compares the robustness and performance of three temporal surveillance techniques using a twofold approach: 1) a unifying statistical analysis to establish their common features and differences, and 2) a benchmarking on influenza-like illnesses (ILI) complaint time series from the Boston area for Harvard Pilgrim Health Care (HPHC). The ILI time series are convenient because "outbreak free" periods can be identified outside of the obvious flu season and used for background data. Furthermore, the daily patient counts outside flu season are low to moderate and therefore more representative of the serious case tracking that would be used to detect the emergence of a dangerous new disease such as SARS or a new virulent flu variety. In the analysis below, we use Poisson models to reflect the sampling with low daily mean. For syndromes with high daily count, such as respiratory infection, applying the Central Limit Theorem and using a normal approximation would have more general applicability. Similar analysis would be developed in this case.

The early stages of an outbreak may be highly stochastic, with additional complexity resulting from the statistical sampling of the infected population by the surveillance system. Not all infected individuals will appear for measurement at a specific emergency department, use an HMO, or shop at a participating pharmacy. Some will go to private practitioners or other hospitals; some will simply not be treated. This statistical sampling will be shown below to be a dominant source of the short-term variation. Both the stochastic nature of an epidemic at its early stages and sampling effects should be considered when developing detection techniques. While the daily rates may not be accurately predictable due to intrinsic, random, day-to-day variations that are difficult to account for, it is conceivable that the underlying dynamics producing the daily variations are governed by transient properties that can be tracked to detect change within a short time period.

The algorithm development and analysis begins with a locally stationary Poisson arrival model for the patients. An outbreak is characterized by an increase in the daily Poisson mean, resulting in stochastic outbreaks. In this paper, we develop detectors for outbreaks with non-decreasing Poisson mean, but the actual random outbreak will rarely be monotonic. This time-varying mean estimate is applied within a detection window of length T, nominally of 3-14 days for early detection. This moving detection window scans the time series in real time, as measurements are available, to detect a current outbreak. The vast majority of the syndromic surveillance literature addresses the detection of step increases in patient levels. We address the more general issue of detecting monotonic increases in mean as characteristic of the early stage of an outbreak following a classic susceptible-exposed-infected-removed model. We address this issue with a set of models of increasing generality. A likelihood ratio provides a common basis for a family of closely related algorithms, which we compare in this paper. Applying a generalized likelihood ratio test (GLRT) to optimize window length results in the G-surveillance statistic [8]; this is itself an adaptation of the widely used scan statistic [9]. G-surveillance uses a step increase in mean as the outbreak model. Three new methods for syndromic surveillance are then considered. A new uniformly most powerful (UMP) test is developed as a matched filter for an entire class of outbreak profiles. This viewpoint, after application of the Central Limit Theorem, is applied in [1] and [2]. Another new approach uses the GLRT formulation in conjunction with a monotonic regression viewpoint for the time varying mean to detect generally increasing (but otherwise stochastic) outbreaks. Finally, a non-negative regression approach is taken to detect outbreaks that are not monotonic. This may especially be useful for non-contagious outbreaks or diseases with long latency periods resulting in clumps of cases at the early epidemic stages.

MASSACHUSETTS SYNDROMIC SURVEILLANCE 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 [3-6]. 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” 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 (ILI), and neurological and were defined in terms of sets of diagnostic codes [7]. A single encounter 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).

METHODS

Modeling the Benchmark Data Set

We develop a straightforward time-dependent Poisson model to describe the time dynamics of X_t, the background noise. Figure 1 shows the original patient count time series. As we will see below, different days of the week have significantly different usage. Consider a catchment of size K_t that constitutes the sample population from which the patient population is selected. The explicit dependence on time allows for a slowly evolving population size. On a particular day t, a member of the catchment population has a certain probability p_t of acquiring a respiratory infection, and thus experiencing a respiratory syndrome. Given that an individual is sick, he or she will then appear at HPHC with a probability q_t. In this case, a simple model for the number of patients arriving on a particular day is binomial

$$X_t \sim B(K_t, p_t q_t) \ . \tag{1}$$

In this context, with $p_t q_t \ll 1$, the binomial is well approximated by a Poisson random variable

$$\begin{aligned} X_t &\sim P(m_t) \\ m_t &= K_t p_t q_t \end{aligned} \tag{2}$$

Several different timescales drive the evolving statistics of X_t. Certainly over periods as long as this study the underlying population size K_t may have changed. Demographic attributes, such as an aging population, may cause p_t to vary, while time of the year and severity of the flu season will also affect p_t. The parameter q_t may be influenced by changes in health care policies and insurance practices, but day of the week variability most significantly affects the probability of arriving at this specific practice given that a respiratory infection has occurred, since office hours are limited on weekends. Further adjustments must also be made for holidays.

Separate measurement of catchment size, illness rates, and daily use are difficult without measurement of many other variables. Therefore, we follow a nonparametric approach, which provides a robust technique applicable to other datasets from different institutions.

The time-varying Poisson mean is estimated through Poisson regression, based on the daily average in a moving time window and adjusted for day of week and holidays. First we capture the recent average level $\bar{m}_t$ using a window of length L. Note that although we use L=14 so that the window contains an equal number of each day of the week, this is not strictly necessary. The Poisson regression model below has a day of the week adjustment that can provide the correction for windows of various sizes.

This average level $\bar{m}_t$ provides a baseline for comparison that does not require a time-of-the-year adjustment. The data set for our benchmarking is not stable enough from year to year to allow simple monthly comparisons. The average is estimated through a low-pass filter with coefficients f_τ

$$\bar{m}_t = \sum_{\tau=0}^{L-1} f_\tau X_{t-\tau-\delta} \tag{3}$$

The guard interval of width δ allows the separation of data used in the mean estimate from data used in the outbreak detection. Choosing a larger guard interval is conceptually simpler, since the baseline mean estimate $\bar{m}_t$ is calculated using a disjoint set of data from the current state X_t. However, larger choices of guard interval introduce larger latencies in the baseline estimate $\bar{m}_t$ and degrade performance. We use $\delta = 0$ below, which leads to the best detector performance. The most obvious choice of filter is a simple block average, with $f_\tau = {}^1\!/\!_L$ for a window of length *L*. Performance, however, is significantly degraded in this case due to the high side lobes associated with this block filter. Since spectral content is the main separating feature between rapid onset of outbreaks (with high-frequency energy) and the slowly varying mean (with low-frequency energy), low filter side lobes are important to overall performance. We address side lobe concerns, while maintaining the interpretation of a time-varying mean estimate, through use of a standard Hamming window for f_τ [10],

$$f_{\tau+1} = 0.54 - 0.46\cos\left(2\pi\frac{\tau}{L-1}\right), \quad \tau = 0,1,\ldots L-1\,. \tag{4}$$

The resulting filter is comparatively long in duration as compared to other low-pass filter designs, but the additional averaging contributes to the robustness of the design.

To accommodate the product form of the expression for $m_t = K_t p_t q_t$, and separate the conditionally independent daily adjustments, we use a Poisson regression model with non-holiday Sundays as the reference:

$$\begin{aligned}\log\left(m_t\right) = \beta_0 + \beta_1\log(\overline{m}_t) + \beta_2 I_{Monday} + \beta_2 I_{Tuesday} + \beta_2 I_{Wednesday} + \beta_2 I_{Thursday} + \beta_2 I_{Friday}\cdots \\ + \beta_2 I_{Saturday} + \beta_2 I_{holiday}\end{aligned} \tag{5}$$

Note, as expected, that the coefficient for $\log(\overline{m}(t))$ is close to one, as expected. Using $\log(\overline{m}(t))$ as an offset parameter (fixing its coefficient at 1) only slightly increases the modeling error. Terms adjusting for months and seasons were considered, but provided minimal benefit.

Applying the Poisson Model

We can now describe our model and hypothesis test for early detection, based on a small window of data of length T. We use the notation $\lambda_{0s}, s = 1,2,\ldots T$ to represent the mean of the Poisson process under the null hypothesis, when no outbreak is present, from our estimator (3-4) with adjustments from (5) . In the benchmarking with HPHC data, we will estimate λ_{0s} with $\widehat{m}_s$ sampled at the appropriate time. Here we demonstrate T=7 and L=14, so the reduced coefficients of a Hamming window reduce the bias in $\widehat{m}_s$ during the earliest stage of an outbreak. We also add a short guard interval $\delta = 7$. Under the null hypothesis, we model our measurement with

$$P((X_1, X_2, \ldots X_T) = (x_1, x_2, \ldots x_T) \mid \mathrm{H}_0) = \prod_{t=1}^{T} \frac{e^{-\lambda_{0t}} \lambda_{0t}{}^{x_t}}{x_t\,!} \,. \qquad (6)$$

Under the alternative hypothesis, for outbreak mean o_t, we have $\lambda_{1t} = \lambda_{0t} + o_t$ and generalized likelihood ratio (GLR)

$$P((X_1, X_2, \ldots X_T) = (x_1, x_2, \ldots x_T) \mid \mathrm{H}_1) = \prod_{t=1}^{T} \frac{e^{-\lambda_{1t}} \lambda_{1t}{}^{x_t}}{x_t\,!}$$

$$\mathrm{GLR} = \underset{o_t \,\in\, \text{class } \mathrm{C_d}}{\arg\max} \prod_{s=1}^{T} \frac{\dfrac{e^{-\lambda_{0s} - o_s} (\lambda_{0s} + o_s)^{Xs}}{X_s\,!}}{\dfrac{e^{-\lambda_{0s}} \lambda_{0s}{}^{X_s}}{X_s\,!}} \,. \qquad (7)$$

Here the outbreak itself is not deterministic, to reflect realistic sampling of the catchment. Only the mean disease level in the catchment increases, with a Poisson model for measurement. Generalized likelihood ratios, with different interpretations, are then used to build different tests, through choice of different classes $\mathrm{C_d}$. Each of the classes is discussed below.

The G-surveillance Statistic

The G-surveillance statistic looks for the best window of length k=1,2,…,T. Use of a generalized likelihood ratio test optimizes window length, leading to a time-domain version of the widely used scan statistic [8,9]. The sum of patients counts over a window of length *k* is also Poisson with

$$\sum_{s=T-k+1}^{T} X_s \sim Po\left(\sum_{s=T-k+1}^{T} \lambda_{0\,s} \right) \qquad (8)$$

under H_0 and

$$\sum_{s=T-k+1}^{T} X_s \sim Po\left(\sum_{s=T-k+1}^{T} \lambda_{1\,s} \right)$$

under H_1 . Note that the maximum likelihood estimate of $\sum \lambda_{1s}$ is $\sum X_s$, which is optimal for outbreaks shaped like step increases. We then use this MLE for the unknown total mean under Ha and have the G-surveillance statistic with a best estimate for window length

$$\tilde{k} = \arg\max_{k=1,2,...,T} \quad e^{\left(\sum_{s=T-k+1}^{T} \lambda_{0\ s}\right)-\left(\sum_{s=T-k+1}^{T} X_s\right)} \left(\frac{\displaystyle\sum_{s=T-k+1}^{T} X_s}{\displaystyle\sum_{s=T-k+1}^{T} \lambda_{0\ s}}\right)^{\left(\sum_{s=T-k+1}^{T} X_s\right)} \qquad (9)$$

and the test is then, for threshold τ ,

$$\mathrm{L_{GS}}\left(X_1, X_2, ..., X_T\right) = \log\left(e^{\left(\sum_{s=T-\tilde{k}+1}^{T} \lambda_{0\ s}\right)-\left(\sum_{s=T-\tilde{k}+1}^{T} X_s\right)} \left(\frac{\displaystyle\sum_{s=T-\tilde{k}+1}^{T} X_s}{\displaystyle\sum_{s=T-\tilde{k}+1}^{T} \lambda_{0\ s}}\right)^{\left(\sum_{s=T-\tilde{k}+1}^{T} X_s\right)} \right) \quad . \quad (10)$$

$$= \sum_{s=T-\tilde{k}+1}^{T} (\lambda_{0\ s} - X_s) \quad - \quad \left(\sum_{s=T-\tilde{k}+1}^{T} X_s\right)\left(\ln\left(\sum_{s=T-\tilde{k}+1}^{T} X_s\right) - \ln\left(\sum_{s=T-\tilde{k}+1}^{T} \lambda_{0\ s}\right)\right) \begin{matrix} \mathrm{H_1} \\ > \\ < \\ \mathrm{H_0} \end{matrix} \quad \tau$$

Uniformly Most Powerful Test

A uniformly most powerful test can be developed as a matched filter for an entire class of outbreak profiles with shape f_s . This viewpoint, after application of the Central Limit Theorem, is applied in [1] and [2]. Consider a fixed time window and

$$\begin{aligned} &\mathrm{H_0}: \quad X_s \sim \mathrm{Po}\left(\lambda_{0s}\right), \qquad s = ..., 0, 1, 2, ...T \\ &\mathrm{H_1}: \quad X_s \sim \begin{cases} \mathrm{Po}\left(\lambda_{0s}\right) & s \le 0 \\ \mathrm{Po}\left(\lambda_{0s}\, e^{\gamma f_s}\right) & s > 0 \end{cases} \qquad . \end{aligned} \qquad (11)$$

Different choices of profile shape o_s capture different phenomena. For example, a constant profile results in a step function, which might be typical of a noncontagious outbreak source. A linear shape describes an exponential increase, as typical in the early stages of a contagious outbreak. The resulting log likelihood ratio test is then

$$\begin{aligned}\mathrm{L}\left(X_1, X_2, ..., X_T\right) &= \ln \prod_{s=1}^{T} \frac{\dfrac{e^{-\lambda_{0s} e^{\gamma f_s}} (\lambda_{0s} e^{\gamma f_s})^{Xs}}{X_s!}}{\dfrac{e^{-\lambda_{0s}} \lambda_{0s}{}^{X_s}}{X_s!}} \\ &= \ln \prod_{s=1}^{T} e^{\lambda_{0s} - \lambda_{0s} e^{\gamma f_s}} (e^{\gamma f_s})^{Xs} \qquad . \\ &= \sum_{s=1}^{T} (\lambda_{0s} - \lambda_{0s} e^{\gamma f_s}) + \gamma \sum_{s=1}^{T} X_s f_s \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \tau \end{aligned} \qquad (12)$$

After identifying

$$\tau' = \frac{1}{\gamma} \left(\tau - \sum_{s=1}^{T} (\lambda_{0s} - \lambda_{0s} e^{\gamma f_s}) \right) \qquad (13)$$

we have the matched filter

$$L_{UMP}\left(X_1, X_2, ..., X_T\right) = \sum_{s=1}^{T} X_s f_s \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \tau' . \qquad (14)$$

Unless both T and λ_{0s} are small, a normal approximation may be developed via the Central Limit Theorem to describe performance in terms of matched filters and signal-to-noise ratio.

Note that this test statistic does not require knowledge of γ to control the false alarm rate. Then the test is uniformly most powerful in the exponential rate γ, so no a priori knowledge of the rate is needed to develop a constant false alarm rate detector. A single detector is used for all exponentially increasing outbreaks, regardless of the exponential rate, with a fixed false alarm rate. Of course, the test power will be dependent on the exponential rate.

Monotonic Regression

Both G-surveillance and the uniformly most powerful test are based on an underlying signal shape o_t. Failure to match the shape will lead to reduced performance. A monotonic regression viewpoint for the time-varying mean adapts to generally increasing (but otherwise stochastic) outbreaks. In application, these growing outbreaks may provide the greatest immediate threat to public health. Applying a monotonic regression viewpoint to the generalized likelihood ratio results in

$$\begin{aligned}\tilde{o} &= \underset{0 \le o_1 \le o_2 \le \ldots \le o_T}{\arg\max} \prod_{s=1}^{T} \frac{\dfrac{e^{-\lambda_{0s} - o_s}(\lambda_{0s} + o_s)^{Xs}}{X_s!}}{\dfrac{e^{-\lambda_{0s}}\lambda_{0s}{}^{X_s}}{X_s!}} \\ &= \underset{0 \le o_1 \le o_2 \le \ldots \le o_T}{\arg\max} \prod_{s=1}^{T} e^{-o_s}\left(\frac{\lambda_{0s} + o_s}{\lambda_{0s}}\right)^{X_s} \qquad . \\ &= \underset{0 \le o_1 \le o_2 \le \ldots \le o_T}{\arg\max} -\sum_{s=1}^{T} o_s + \sum_{s=1}^{T} X_s\left(\ln(\lambda_{0s} + o_s) - \ln\lambda_{0s}\right)\end{aligned} \tag{15}$$

Many techniques are available to solve the monotonic regression problem. Because of the small window size, we avoid numerical issues with a direct search over the constraint combinations in the argmax .

$$\mathrm{L}_{\mathrm{MR}}\left(X_1, X_2, \ldots, X_T\right) = -\sum_{s=1}^{T} o_s + \sum_{s=1}^{T} X_s\left(\ln(\lambda_{0s} + o_s) - \ln\lambda_{0s}\right) \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \tau \tag{16}$$

Note again that the outbreak itself is stochastic, with the profile o_s describing and increase in disease in the catchment and Poisson sampling applied. The maximum likelihood outbreak shape $\tilde{o}_s$ also provides an inference on the characteristics of the outbreak.

Non-negative Regression

A non-negative regression approach is taken to detect outbreaks that are not monotonic. This may especially be useful for non-contagious outbreaks or diseases with long latency periods resulting in clumps of cases at the early epidemic stages. The maximum likelihood estimate of outbreak shape is then $\max(X_s, \lambda_{0\,s})$ with

$$\begin{aligned}
\tilde{o} &= \underset{0 \le o_s}{\arg\max} \prod_{s=1}^{T} \frac{\dfrac{e^{-\lambda_{0s} - o_s}(\lambda_{0s} + o_s)^{Xs}}{X_s!}}{\dfrac{e^{-\lambda_{0\,s}} \lambda_{0\,s}{}^{X_s}}{X_s!}} \\
&= \underset{0 \le o_s}{\arg\max} \prod_{s=1}^{T} e^{-o_s} \left(\frac{\lambda_{0\,s} + o_s}{\lambda_{0\,s}} \right)^{X_s} \qquad (17) \\
&= \underset{0 \le o_s}{\arg\max} - \sum_{s=1}^{T} o_s + \sum_{s=1}^{T} X_s \left(\ln(\lambda_{0\,s} + o_s) - \ln \lambda_{0\,s} \right)
\end{aligned}$$

and

$$\tilde{o}_s = \max(X_s, \lambda_{0\,s}) \,.$$

For the generalized log likelihood detector,

$$\mathrm{L}_{\mathrm{NNR}}\left(X_1, X_2, ..., X_T\right) = \sum_{s=1}^{T} (\lambda_{0\,s} - \max(X_s, \lambda_{0\,s})) + \sum_{s=1}^{T} X_s \left(\ln\left(\max(X_s, \lambda_{0\,s})\right) - \ln \lambda_{0\,s} \right) \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \tau \,. \qquad (18)$$

This detector provides the most general detection strategy but, as seen below, may have less power for monotonic outbreaks.

The Optimal Test

The theoretically optimal test provides the benchmark for comparison of the various techniques. This optimal test is based on two unrealistic assumptions: the background process mean $\lambda_{0\,s}$ and outbreak process mean $\lambda_{1\,s}$ are exactly known a priori, and the resulting measurements are exactly Poisson processes, independent at each time. Obviously neither assumption is accurate in practice, but the optimal detector does provide an upper performance bound.

Instead of generating a test for a fixed mean, we instead map our actual data set into a Poisson process, independent at each time, with known time varying means $\lambda_{0\,s}$ and $\lambda_{1\,s}$. By using actual experimental time series to specify a candidate $\hat{\lambda}_{0\,s}$, based on our Poisson regression, we provide some level of realism while applying a theoretically optimal test. The mean during outbreak is also considered known with $\lambda_{1\,s} = \hat{\lambda}_{0\,s} + o_s$. In this case, we can specify and apply the optimal test with a log likelihood ratio and

$$
\begin{aligned}
\mathrm{L}_{\mathrm{OPT}}\left(X_1, X_2, ..., X_T\right) &= \ln \prod_{s=1}^{T} \frac{\dfrac{e^{-\hat{\lambda}_{0\,s}-o_s}\,(\hat{\lambda}_{0\,s}+o_s\,)^{Xs}}{X_s\,!}}{\dfrac{e^{-\hat{\lambda}_{0\,s}}\,\hat{\lambda}_{0\,s}{}^{X_s}}{X_s\,!}} \\
&= \ln \prod_{s=1}^{T} e^{-o_s} \left(\frac{\hat{\lambda}_{0\,s}+o_s}{\hat{\lambda}_{0\,s}} \right)^{Xs} \qquad . \qquad (19) \\
&= -\sum_{s=1}^{T} o_s + \sum_{s=1}^{T} X_s \ln\left(\frac{\hat{\lambda}_{0\,s}+o_s}{\hat{\lambda}_{0\,s}} \right) \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \tau
\end{aligned}
$$

Test Thresholds

Each of the tests above require calculation of a threshold for the test statistic. Due to the complex form of the tests, no simple analytical solution is available. Instead, we use our null hypothesis model based on the time-varying Poisson process resulting from our mean estimation procedure, λ_{0s}. Then our model for the measured vector in the window is

$$P((X_1, X_2, ...X_T) = (x_1, x_2, ...x_T) \,|\, \mathrm{H}_0) = \prod_{t=1}^{T} \frac{e^{-\lambda_{0t}} {\lambda_{0t}}^{x_t}}{x_t!} . \tag{20}$$

This model is used with repeated sampling to generate test thresholds for each of our statistics for a targeted test size.

The actual test data set only approximately meets this model under H_0. To assess performance, the actual time series is then used to estimate actual detection and false alarm probabilities. Due to the approximations made in the model for the threshold, and in particular the assumption of independence, the actual test size is slightly larger than the target. Of course, this can be accommodated in the design. To facilitate detection comparison, we estimate actual detection and false alarm probabilities below.

RESULTS—SENSITIVITY AND SPECIFICITY

Our test time series clearly contains outbreaks during the peak of flu season. In our performance valuations for detector performance, we exclude these periods of obvious outbreak using a simple threshold. If $\hat{m}_t > 20$ at any time during the week used in the detector, we strongly suspect an outbreak was present in the initial data set and exclude that time from analysis. With this requirement, 89.75% of the data set was used and 10.25% was excluded. Successful detections during these highly spiked periods are difficult to interpret, since the nominally H0 testing time series unfortunately contains outbreaks. This exclusion allows us to assess performance during periods with a lower likelihood of already containing an outbreak.

To evaluate the relative detector performance, we spiked the ILI time series with a random outbreak following a Poisson distribution λ_{1s}, independent in time, using a one-week-long analysis window. Two different outbreak shapes are considered: an exponentially increasing outbreak typical of contagious disease, and a step increase. To evaluate timeliness, detection tests were run with only one day of outbreak in the test window, two days of outbreak in the test window, and so on until the one-week analysis window contained only outbreak day. Since the uniformly most powerful test has the possibility of model/outbreak disparity, the mismatch case was also analyzed for UMP. Outbreaks were spiked and actual false alarm and detection probabilities were tabulated in Table 2. The optimal test provides an upper bound on performance, using the assumption that the null hypothesis model is accurate.

DISCUSSION

Model Validation: Poisson Distribution of Patient Counts

We first confirm two properties of the distribution: the Poisson distribution, and the decorrelation between time samples. We demonstrate our approach on our HPHC dataset, but the reader should note that our technique captures the population health seeking behavior and is generally applicable to different datasets from different institutions. The Poisson assumption itself is not critical, especially for larger daily counts, due to the Central Limit Theorem, but the assumption that the time-varying mean approximates the time varying variance is critical. Our Poisson regression approach in equation (5) leads to a time-varying estimate of the mean $\hat{m}(t)$, based on levels during the last two weeks, day of the week, and holiday status.

Testing this approximation is complicated because the mean m_t changes too rapidly to allow enough averaging for a high-quality estimate. Instead, we follow a diagnostic viewpoint that is commonly applied in logistic regression. We sort the values of X_t and $\hat{m}_t$ in ascending order based on $\hat{m}_t$, resulting in $\tilde{X}_s$. In this sorted data set, samples close together in time have nearly the same estimated mean $\hat{m}_s$. Since the standard deviation provides the natural scale for error probabilities, we then estimate $std\left(\tilde{X}_s\right)$ using the ensemble of time points with nearest values of $\widehat{m}_s$. We test this approximation on our historic time series. Figure 2 shows the resulting $\hat{m}_s{}^{1/2}$ and estimated $std\left(\tilde{X}_s\right)$, using the sorted values. Note that $\hat{m}_s{}^{1/2}$ is a reasonable approximation of $std\left(\tilde{X}_s\right)$, except for large values of $\hat{m}_s{}^{1/2}$. This approximation is not perfect and would certainly fail any statistical test of fit. In particular, the match is weakest during periods of highest mean level. These high mean periods represent time-localized outbreaks in the original data set, so our low-pass filter estimate for outbreak-free levels should be inaccurate during these times. However, the approximation is based on a simple underlying model—the Poisson sampling of patients—and as such is highly robust and widely applicable. Use of this approximation allows the development of early surveillance approaches without the expensive, arduous, and often impossible task of collecting many years of syndromic-specific data for each location.

Model Validation: Prewhitening

In the usual development of matched filter designs, we would derive a prewhitening filter to specifically match the spectrum of the data set. This approach requires large historic data sets, so we avoid it here. Instead, we demonstrate that our simple locally stationary Poisson model acts as a prewhitening filter.

Under a no-outbreak condition, we have developed an approximate distribution for the marginal distribution of X_t. We now investigate time dependence using the normalized autocorrelation

$$\frac{E\left(\left(X_t-\hat{m}_t\right)\left(X_{t+\tau}-\hat{m}_{t+\tau}\right)\right)}{E\left(\left(X_t-\hat{m}_t\right)^2\right)} \tag{21}$$

for our historic time series as shown in Figure 3. Note that the time series for $X_t-\hat{m}_t$ substantially decorrelates in a single day. We can model X_t as uncorrelated from time sample to time sample. The Central Limit Theorem would then suggest that the Poisson time series is independent, as required in the theoretical model. This is only approximately true for periods of low to moderate daily count.

Comparing the Detectors

Table 2 provides a comparison of the power and timeliness of all four techniques, benchmarked with the HPHC time series. In particular, the tradeoff between timeliness, robustness, and maximum detection power has been established in a wide number of settings and illustrated with the example in Table 2. An actual probability of false alarm of 0.02 was used for analysis, which would correspond to approximately one false alarm event per year, given the approximately one week correlation period in the detector. Obviously, detection was nearly impossible in the first day and progressively easier as the outbreak filled the week-long analysis window. None of the approximate detectors reached the performance of the optimal test, as expected since the time series only approximately fit the model. However, at the seventh day of the outbreak, performance of the correctly matched UMP test was comparable to the optimal test, despite limitations in modeling. In this example, and the many other numerical experiments we undertook, the monotonic regression consistently performed well. The monotonic regression even performed well against the uniformly most powerful test during UMP filter/outbreak match, despite the fact that monotonic regression is much less parametric.

The uniformly most powerful test performed well, given that it tested for the presence of an entire class of outbreaks. This is particularly attractive for testing for exponential growth families, since the growth parameter would rarely be known a priori. However, model filter/outbreak shape mismatch does cause some deterioration and reduces the UMP performance to that of the monotonic regression approach.

Performance of the G-surveillance statistic, however, was generally inferior. In retrospect, this is not surprising. The spatial scan statistic must identify the geographical location of an outbreak to detect time-space localized outbreaks. The generalized likelihood test for this location is critical. In a time scan setting, the time windows are all nested inside the one week analysis window. The maximum operation is in some sense superfluous, since a fixed one-week window can still be used to capture the full outbreak energy. In fact, a fixed constant window did outperform G-surveillance, especially when the transient filled the entire outbreak window.

The non-negative regression statistic performed consistently below the other test statistics. This is somewhat surprising, since monotonic regression provided consistently good performance. Non-negative regression does seek a quite general signal structure, making outbreaks more difficult to separate from the underlying noise. This suggests that the more general problem of identifying any outbreak, regardless of shape, is significantly more difficult than early detection of increasing outbreaks.

CONCLUSIONS

Comparing and selecting temporal surveillance techniques are complicated tasks because of the wide variety of techniques in the literature. In this paper, we demonstrate an approach for this process based on common analytical development and benchmarking with common data sets. Poisson regression is applied to develop a time-varying model of the number of visits classified as influenza-like illness to the Harvard Pilgrim Health Care system in the Boston area. Four different outbreak detection approaches are contrasted using this common data set, and the optimal detector performance is provided as a benchmark.

A highly nonparametric approach, based on monotonic regression, provided the best overall performance and most detection robustness for early detection of stochastic, increasing outbreaks. In situations where an exponentially growing outbreak is the target, then the uniformly most powerful test delivers nearly optimal performance. However, more work remains to be done. Our approach here, with use of common data sets and analytical assumptions, provides a uniform approach for evaluating and comparing surveillance techniques. While additional work in developing effective surveillance techniques are needed, it is likely that different detection techniques may perform differently under various outbreak conditions or datasets. For example, a monotonic regression algorithm may be better suited for capturing the early transmission dynamics of contagious disease while others respond better to retrospective detection of recurring noninfectious disease processes. Finally, some techniques may be region-specific and sensitive to localized clustering of disease incidents in time and space while others detect elevated numbers across an entire area. Relative advantages and disadvantages of different surveillance techniques cannot be systematically addressed until a uniform evaluation approach is adopted.

REFERENCES

[1] Reis BY, Pagano M, Mandl KD, Using temporal context to improve surveillance, PNAS. 2003; 100(4):1961-1965.

[2] Dunyak, J., and Mandl, K., Applying Emergency Department Complaint Codes to Detect Disease Outbreaks. Submitted to IEEE Trans. On Information Technology in Biomedicine. Under review.

[3] Lazarus R, Kleinman K, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes: the example of lower respiratory infection. BioMed Central Public Health 2001;1:1–9.

[4] 2. Lazarus R, Kleinman K, Dashevsky I, Adams C, Kludt P, DeMaria Jr. A, Platt R. Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events. Emerg Infect Dis 2002;8:753–760.

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

[6] Yih WK, Caldwell B, Harmon R, Kleinman K, Lazarus R, Nelson A, Nordin J, Rehm B, Richter B, Ritz-woller D, Sherwood E, Platt R. The National Bioterrorism Syndromic Surveillance Demonstration Pro-gram. In: Syndromic Surveillance: Reports from a National Conference, 2003. Morbidity and Mortality Weekly Report 2004;53 (suppl):43-46.

[7] CDC. Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Available at http://www.bt.cdc.gov/surveillance/syndromedef/index.asp.

[8] Wallenstein, S. and Naus, J. Scan Statistic for Temporal Surveillance for Biological Terrorism. Morbidity and Mortality Weekly Report. 2004; v. 53, p. 74-78.

[9] Naus J. The distribution of the size of the maximum cluster of points on a line. J Am Stat Assoc 1965;60:532--38. 5.

[10] Van Trees HL. Detection, Estimation, and Modulation Theory, Part I. Wiley-Interscience, 2001.

terms	constant term	$\log(\bar{m}(t))$	I_{Monday}	$I_{Tuesday}$	$I_{Wednesday}$	$I_{Thursday}$	I_{Friday}	$I_{Saturday}$	$I_{Holiday}$
β	-1.0867	0.9587	0.8663	0.7257	0.6533	0.6573	0.5596	0.0096	-0.6230

Table 1: Coefficients in the Poisson regression model, using a Hamming window of length L=14

Table 2a: Detection probabilities for an exponentially increasing outbreak.

outbreak length/ shape λ_{1s}	1 day λ_{0s} *[0 0 0 ... 0 0 0 0.10]	3 day λ_{0s} *[0 0 0 ... 0 0.10 0.22 0.35]	5 day λ_{0s} *[0 0 0.10 ... 0.22 0.35 0.49 0.64]	7 day λ_{0s} *[0.10 0.22 0.35 ... 0.49 0.64 0.81 1.00]
G-surveillance	0.02	0.05	0.22	0.70
UMP for exponential families	0.03	0.09	0.41	0.85
monotonic regression	0.03	0.10	0.34	0.76
non-negative regression	0.02	0.07	0.25	0.69
optimal	0.03	0.12	0.39	0.85

outbreak length/ shape λ_{1s}	1 day λ_{0s} *[0 0 0 ... 0 0 0 0.60]	3 day λ_{0s} *[0 0 0 ... 0 0.60 0.60 0.60]	5 day λ_{0s} *[0 0 0.60 ... 0.60 0.60 0.60 0.60]	7 day λ_{0s} *[0.60 0.60 0.60 ... 0.60 0.60 0.60 0.60]
G-surveillance	0.09	0.35	0.55	0.72
UMP for constantl families	0.05	0.28	0.64	0.87
monotonic regression	0.17	0.50	0.65	0.79
non-negative regression	0.11	0.37	0.59	0.75
optimal	0.29	0.57	0.74	0.87

.

Table 2b: Detection probabilities for a constant outbreak.

outbreak length/ UMP shape Filter Actual outbreak	1 day [1 2 3 4 5 6 7] $\lambda_{0\,s}$ *[0 0 0 ... 0 0 0 0.60]	3 day [1 2 3 4 5 6 7] $\lambda_{0\,s}$ *[0 0 0 ... 0 0.60 0.60 0.60]	5 day [1 2 3 4 5 6 7] $\lambda_{0\,s}$ *[0 0 0.60 ... 0.60 0.60 0.60 0.60]	7 day [1 2 3 4 5 6 7] $\lambda_{0\,s}$ *[0.60 0.60 0.60 ... 0.60 0.60 0.60 0.60]
UMP for exponential families	0.10	0.49	0.75	0.79
outbreak length/ UMP shape Filter Actual outbreak	1 day [1 1 1 1 1 1 1] $\lambda_{0\,s}$ *[0 0 0 ... 0 0 0 0.10]	3 day [1 1 1 1 1 1 1] $\lambda_{0\,s}$ *[0 0 0 ... 0 0.10 0.22 0.35]	5 day [1 1 1 1 1 1 1] $\lambda_{0\,s}$ *[0 0 0.10 ... 0.22 0.35 0.49 0.64]	7 day [1 1 1 1 1 1 1] $\lambda_{0\,s}$ *[0.10 0.22 0.35 ... 0.49 0.64 0.81 1.00]
UMP for constant families	0.02	0.05	0.26	0.74

Table 2c: UMP detection probabilities during model mismatch.

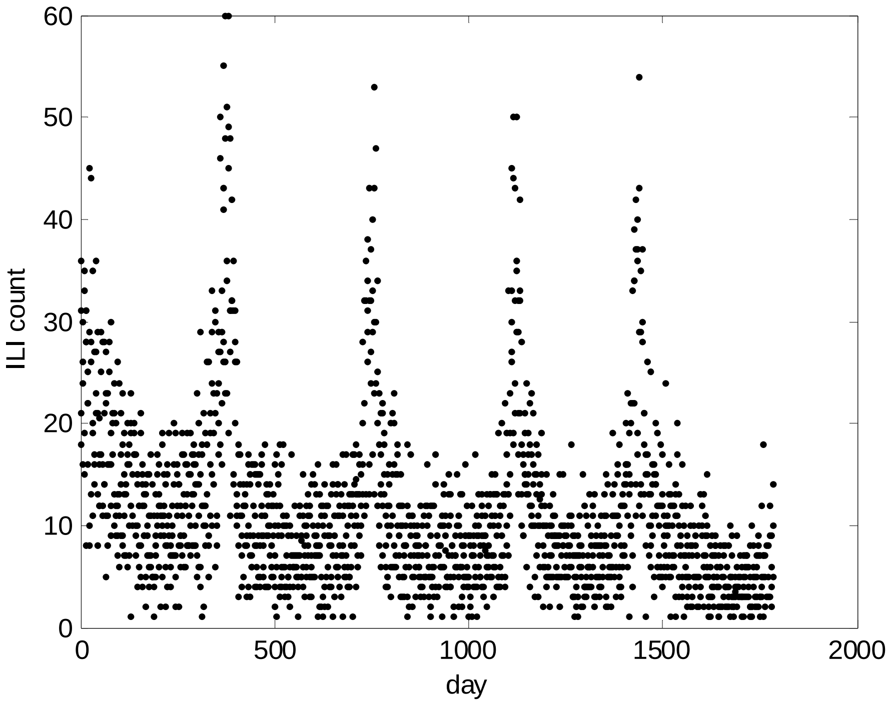

Figure 1: ILI daily patient counts from Harvard Pilgrim Health Care. Day-of-week variability contributes substantially to the overall variability.

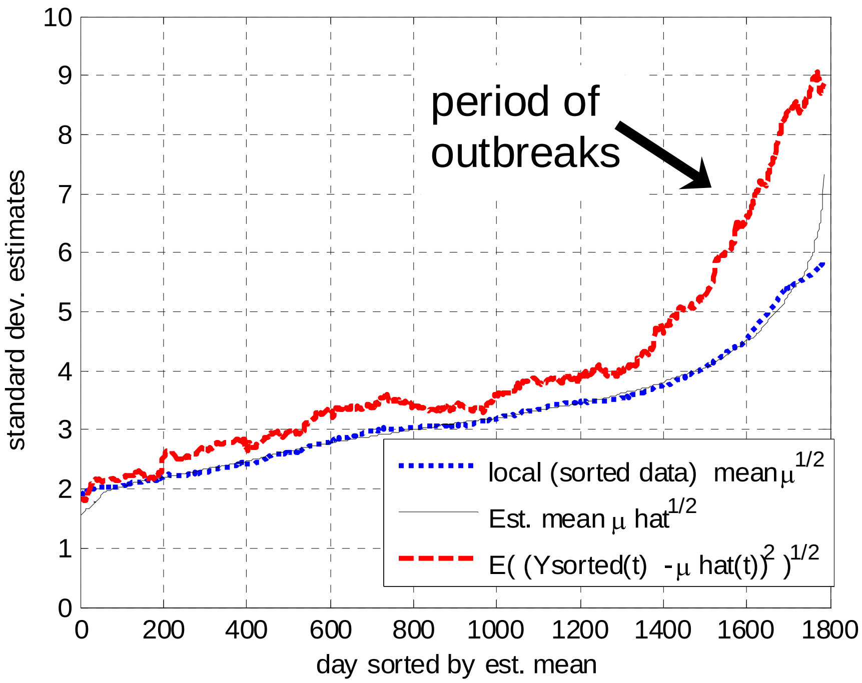

Figure 2: Sorting the days based on estimated Poisson mean $\hat{m}_s$, allows model comparison of square root of means and standard deviations for similar days. Using a sliding window of 14 days, the square root of the estimated mean $\hat{m}_s^{1/2}$ and sorted mean are very close. They also well approximate the standard deviation, as expected for Poisson processes.

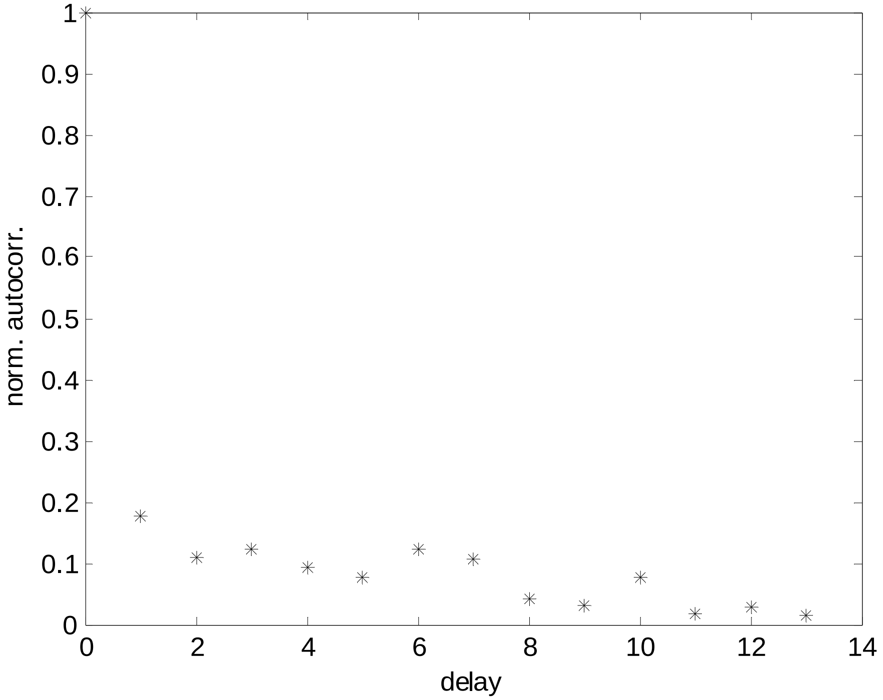

Figure 3: Normalized autocorrelation function. Note that the process substantially decorrelates in a single day, supporting an independent approximation.