Developments in the Roles, Features, and Evaluation of Alerting Algorithms for Disease Outbreak Monitoring

Howard S. Burkom, Yevgeniy Elbert, Steven F. Magruder, Amir H. Najmi, William Peter, and Michael W. Thompson

Automated systems for public health surveillance have evolved over the past several years as national and local institutions have been learning the most effective ways to share and apply population health information for outbreak investigation and tracking. The changes have included developments in algorithmic alerting methodology. This article presents research efforts at The Johns Hopkins University Applied Physics Laboratory for advancing this methodology. The analytic methods presented cover outcome variable selection, background estimation, determination of anomalies for alerting, and practical evaluation of detection performance. The methods and measures are adapted from information theory, signal processing, financial forecasting, and radar engineering for effective use in the biosurveillance data environment. Examples are restricted to univariate algorithms for daily time series of syndromic data, with discussion of future generalization and enhancement.

EVOLVING ROLE OF ALERTING ALGORITHMS IN HEALTH SURVEILLANCE

Research has been ongoing to adapt and implement health surveillance algorithms for decades and has accelerated since the late 1990s because of concern over the threat of a clandestine bioterrorist attack.[1] Methods have been imported from many disciplines in which prospective signal detection has been applied, and some experience has been gained. Successes have been limited because of the rarity of large-scale outbreaks, the lack of documentation of community-level ones, and the lack of consensus over which data effects are outbreak signals and which should be considered background noise. However, the application continues to grow more important and more difficult because of advances in public health informatics and the increased perception of natural disease threats, such as pandemic influenza, along with bioterrorism concerns. More and

increasingly complex data streams are available to health monitors, methodologies for monitoring them need to keep pace, and automated algorithms are needed to direct the attention of investigators to potential problems amid this ocean of information.

Expert knowledge—a combination of local background data experience, medical knowledge, and understanding of population behavior—has not proven sufficient to bridge the gap between statistical and epidemiological significance of algorithmic alerts. The use of automated biosurveillance systems has thus produced false-alarm problems,[2] and most of the reported benefit of these systems has been in combination with other surveillance tools, as in corroboration of medical suspicions and a broadening of situational awareness. Current algorithm research is addressing how to improve detection performance for greater specificity and broader utility. Figure 1 sketches the roles of alerting algorithms in a biosurveillance system, with elements of routine operations on the right and design and evaluation functions on the left.

The algorithmic challenges are as follows:

- To capture the best datasets for making effects of an outbreak stand out as much as possible from usual data behavior.
- To filter the records to produce time series for additional outbreak discrimination, i.e., for maximizing the signal-to-noise ratio.[3]
- To develop and apply robust, sensitive algorithms tuned to these time series to identify the signal as early as possible. Key components of these algorithms are prediction and anomaly determination.[4] Prediction is important because unusual data behavior cannot be recognized without estimates of usual behavior. Systematic time series effects such as seasonal cycles, day-of-week usage patterns, and regular clinic closings are common features in syndromic time series. Various forecasting approaches have been applied to remove expected trends and outliers so that the control charts could be applied to the forecast residuals to determine when to alert.

To provide perspective on development of an entire biosurveillance system, the right half of Fig. 1 shows the additional challenges of data acquisition, cleaning, and transfer at the top and the challenge of useful visualization for human interpretation at the bottom. The design and maintenance of the data chain are especially important; the utility of algorithms, output products, and interfaces depends on the prompt availability of secure data at the required level of detail.

The remaining sections of this article discuss our recent efforts on algorithmic subtasks depicted in Fig. 1 and applied for the ongoing development of the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE). For more

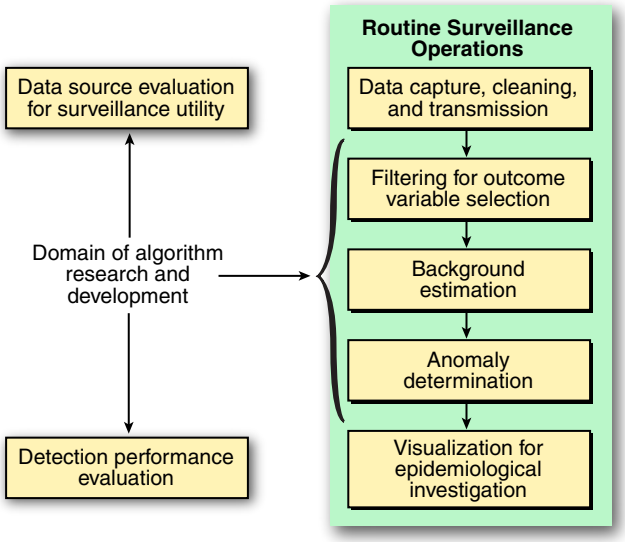

Figure 1. Components of biosurveillance systems, including the routine algorithmic alerting process.

background information on ESSENCE, see the article on biosurveillance research and policy tradeoffs by Burkom et al. elsewhere in this issue. These sections present

- The challenge of outcome variable selection, featuring a discussion of mutual information, its application to a surveillance time series, and a reference series to determine whether their quotient, such as the division of daily syndromic counts by total daily facility visits to estimate a population-level rate, can reduce the signal-to-noise ratio
- An adaptive, recursive least-squares (RLS) algorithm tailored for univariate, city-level syndromic time series with applications for multivariate analysis
- Improvements in implemented algorithms previously reported and a strategy for replacing them with a more data-adaptive approach

Although detailed material in this article is restricted to univariate algorithms, multivariate alerting methods also are conservatively applied in ESSENCE systems, and their broader application is an active research area whose niche in surveillance practice is still to be determined. The final section discusses research challenges of current interest for both univariate and multivariate detection methods.

USE OF MUTUAL INFORMATION FOR IMPROVED SIGNAL-TO-NOISE RATIO

Monitoring in the Absence of a Static Baseline

Most surveillance systems are vulnerable to dramatic and unpredictable shifts in the monitored health care data[3] resulting from changes in data collection methods, diagnosis coding, participating clinical facilities,

insurance eligibility, and similar factors. Reis et al.[4] have noted that anomaly detection methods in most surveillance systems are not robust to shifts in health care utilization because they cannot adjust quickly to changing baselines, so that utilization shifts may trigger false alarms. As a result, the effects of public health crises and major public events may undermine health surveillance systems at the very times they are needed most.[4] Baseline changes may trigger irrelevant alarms and mask signals caused by population health events of concern. For example, a sudden jump in influenza-like illness diagnoses in a population might be caused by an increase in the size or perhaps insurance eligibility of the monitored population and not by a real disease outbreak. In addition, the total number of medical facility visits could rise because of the addition of a hospital into the network, but the hospital might not have a clinic treating the syndrome of interest.[3]

Avoiding false alerts in syndromic monitoring by comparing data to a baseline population is made difficult by the lack of a stable data background. True incidence rates, classically measured by the number of new cases in a given time period divided by the population size,[5] can rarely be calculated in syndromic surveillance data because catchment area sizes for the data sources typically are unavailable. These data sources include records of deidentified and filtered clinical encounter records, pharmaceutical prescriptions, or selected over-the-counter remedy purchases. Although it is possible to estimate the population in a particular geographical catchment area, this estimate may not be useful as a denominator for proportion monitoring because it cannot include patient preferences, temporary closures, or sales promotions in pharmacies or diagnostic laboratories or large transient population inflow or outflow caused by long holiday weekends or large public events (political conventions, the SuperBowl, etc.). The number of people served by a military hospital is particularly difficult to estimate because of the transitory nature of military lifestyles.

Syndromic Data Categorization

In biosurveillance, a syndrome grouping is defined as being a filtering of clinical prediagnostic records designed to clarify the data signal resulting from outbreaks of certain disease types.[6] Clinical encounter data compiled and formatted by systems such as ESSENCE[7] are cleaned to remove duplicate, incomplete, or ambiguous entries. Before records are available for analysis, data fields with personal identifiers are removed or altered to comply with privacy requirements while enough is made available for accurate filtering and analysis. The data streams available for alerting algorithms typically consist of these cleaned records and may come from various sources (e.g., syndromic hospital visit counts and school absentee rates).

Challenge of Minimizing False Alerts While Maintaining Signal-to-Noise Ratio

One approach to reducing false alarms in syndromic data streams is to approximate true rates by replacing daily filtered diagnosis counts with the quotient of these counts by some reference series, such as the total of all daily visits. The Centers for Disease Control and Prevention (CDC) uses this approach for seasonal influenza surveillance by monitoring weekly percentages of influenza-like illness submitted by participating sentinel physicians. The rationale for monitoring ratios instead of counts is that general health care utilization shifts and other data features irrelevant to disease surveillance often affect both numerator and denominator and thus are reduced or eliminated in the ratio, thereby lowering the false-alarm rate. Thus, changes in the syndromic data stream (the ratio's numerator, denoted below as the "target" data stream[4]) relative to a changing background (the denominator, denoted the "context" stream) are drawn out, and these changes are presumably more likely to indicate genuine increases in the syndrome group of interest.

An example of the signal-to-noise ratio increase obtained by using proportions instead of counts is illustrated in Fig. 2. Two plots are shown in the upper half of the figure: the red line is a plot of the daily total number of visits recorded in a health care facility, and the blue line shows counts of only those visits classified in the respiratory syndrome. A small simulated outbreak is indicated in the count series. The outbreak signal is much more distinct in the time series of ratios than in the series of unchanged counts.

The purpose of the work described here is to provide the epidemiologist or system designer with a mathematical tool to decide whether to use individual counts or to use proportions, and which proportions, when monitoring daily surveillance data. The approach herein complements the proportional model networks of Reis et al.[4] and may be used to decide which target/context ratios should be monitored in a multisource network, thus reducing the network size and computational complexity.

Technical Approach for Evaluating Ratios for Signal-to-Noise Discrimination

The approach described here to solve the "denominator problem" is to use information-theoretic techniques to determine under what conditions proportions are preferable to counts and whether ratios formed by using a particular reference series are more effective than simple counts for anomaly detection. Given two data streams $X = \{x(t)\}$ and $Y = \{y(t)\}$, their *mutual information* $I(X; Y)$ provides a general measure of their interdependency. Let X have a probability density function (pdf) given by $p(x)$, and let $p(y)$ be the pdf of Y. If their joint

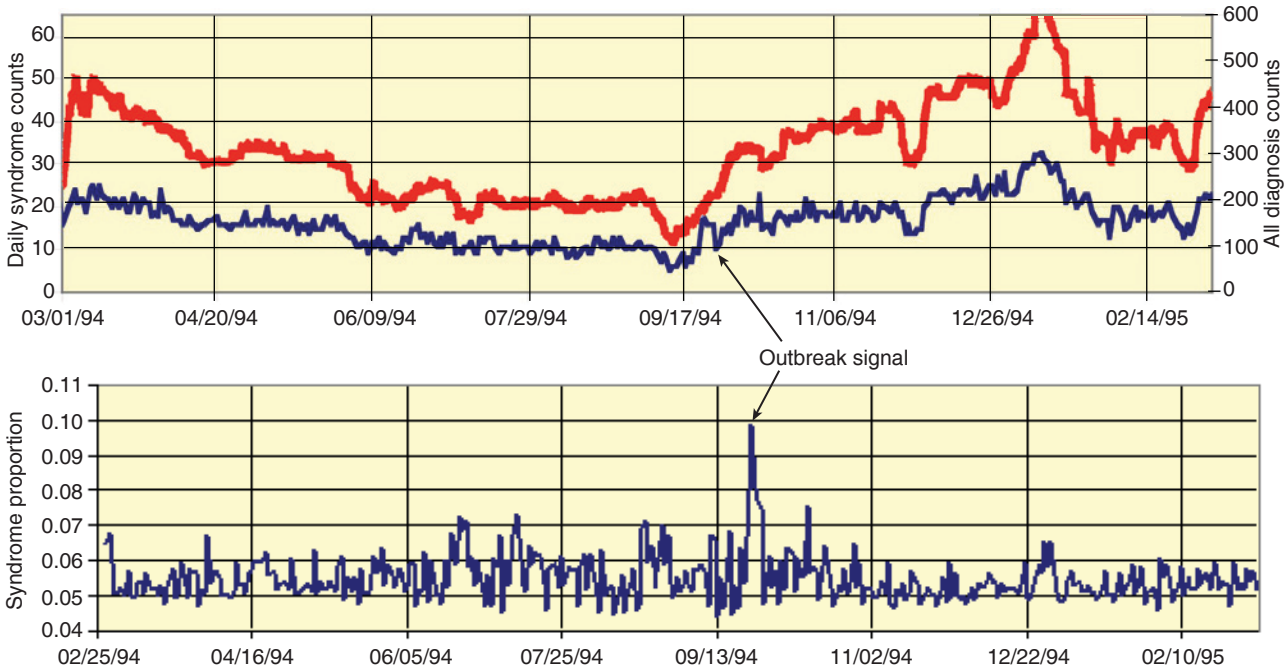

Figure 2. When a denominator variable (or context stream) such as total diagnosis count is available, a proportion sometimes can clarify an outbreak signal.

pdf is denoted by $p(x, y)$, then the mutual information $I(X; Y)$ between X and Y is defined formally as[8]

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \left[\frac{p(x,y)}{p(x)p(y)} \right]. \quad (1)$$

Note that the mutual information also can be written as

$$I(X;Y) = H(X) + H(Y) - H(X,Y)\ , \quad (2)$$

where $H(X)$ is the usual Shannon entropy of a single variable defined by

$$H(X) = -\sum_x p(x) \log p(x) \quad (3)$$

and the joint entropy $H(X, Y)$ is given by[8]

$$H(X,Y) = \sum_x \sum_y p(x,y) \log p(x,y)\,. \quad (4)$$

Mutual information (*MI*) is a general measure of dependency in data: a nonzero *MI* indicates that two data streams X and Y share some kind of dependence, linear or not. We use *MI* to decide which data streams are more appropriate for use as a context stream to a given target stream. In essence, we are designing a noise filter by using a context channel to remove the noise from the target channel. *MI* is a nonlinear generalization of the well-known linear dependency summary statistic, the Pearson linear correlation. The advantages of using *MI* over Pearson correlation as a summary statistic can be seen by the following example. Let a random variable X be uniformly distributed on the interval from 0 to 1, and let $Y = X^2$. Then Y is completely determined by X, so that X and Y are dependent, but their correlation is zero; they are uncorrelated. The mutual information between X and Y, however, is significant.

MI not only provides a general measure of dependency between two variables, but it also has the important feature (for biosurveillance applications) of being robust to missing data values and has been shown to be advantageous in analyzing datasets in which up to 25% of the values are missing.[9] For these reasons, the mutual information metric has been used to analyze dependency in such diverse fields as bioinformatics,[9–11] physics,[12,13] and ecology.[14]

Mutual Information Estimation

Calculating precise values for *MI* is nontrivial, and the accurate estimation of mutual information has been discussed by Kraskov et al.[15] and Steuer et al.[16] The most straightforward technique is to use a histogram-based procedure.[11] In this method, a bivariate histogram is used to approximate the joint pdf of the variables. The use of histograms requires an appropriate choice of binning strategy to obtain an accurate but economical estimate of the joint pdf. Some popular binning strategies are those of Sturges[17] (lower bound on number of bins $k \sim 1 + \log_2(N)$, where N is the data size), Law and Kelton[18] (upper bound on number of bins $k \sim N^{1/2}$), (number of bins $k \sim 1 + \log_2(N)$), and Scott[19] (bin width $h = 3.5s/N^{1/3}$, where s is the sample standard deviation).

Other authors have chosen to circumvent the histogram problem by using different techniques to estimate the joint pdf. These include adaptive binning,[12] k nearest-neighbor statistics,[15] kernel density estimators,[20] and B-spline functions.[10] We have compared some of the different methods of calculating mutual information and found little relative difference in the MI rankings for pairs of variables in a dataset. The absolute magnitudes of $I(x; y)$ for the same pair of variables, however, did change with the particular algorithm employed (depending on the accuracy of the approximation method and on the normalization used).

As seen from Eq. 4, mutual information is bounded above by the sum $H(X) + H(Y)$, so that if this sum is small, $I(X; Y)$ can be small even if X and Y are completely correlated. To obtain a dependency measure whose magnitude is meaningful, some authors normalize MI so that its maximum value is unity if X and Y are completely correlated, similar to the behavior of the correlation coefficient ρ. Normalization techniques include division of $I(X; Y)$ by the arithmetic mean[21] $(H(X) + H(Y))/2$, by the maximal marginal entropy of the considered dataset,[22] or by the geometric mean $(H(X)H(Y))^{1/2}$ of the individual entropies.[23]

The mutual information here was calculated in two ways: the first by using a histogram method with a binning procedure following Priness et al.[11] and the second by using the B-spline method outlined by Daub et al.[10] When using the binning technique, we found it convenient to use an arithmetic mean normalization.[19] The B-spline calculations were performed by using the source code in C++ made freely available by Daub et al.[10] for noncommercial use. The choice of mutual information estimation algorithm did not significantly change the conclusions of our study, so the MI calculations shown below will be those calculated with the B-spline algorithm, whose source code is easily obtained.[10]

Testing Benefits of Mutual Information-Based Discrimination Using Simulation

Based on the discussion above, a series of Monte Carlo simulations were conducted on time series of deidentified, syndromic outpatient visit counts made available for ESSENCE research. The simulations were implemented by introducing simulated lognormal spikes in the daily counts of the health time series to simulate outbreak data effects. In a representative alerting algorithm used for evaluation in this study, series z-score values above a given threshold were considered to be alerts, where a z score was obtained by subtracting the sliding baseline mean from the current count and dividing by the baseline standard deviation. This simple detector was implemented over a large number of simulations with varying thresholds and spike magnitudes to compare the probability of detection in (*i*) daily counts in a single-syndrome data stream and (*ii*) daily counts made up of the ratio of this single-syndrome data stream with other syndromic data streams (including the total daily visit count).

We illustrate typical simulation results using daily counts of rash syndrome visits over a 1402-day period. The time series of these rash syndrome counts is plotted in the upper graph of Fig. 3 and labeled R. A data stream constructed from a Poisson-distributed random time series with mean $\lambda = 50$ over the same time period is shown in the lower graph of Fig. 3 and denoted by P. Monte Carlo simulations then were performed on (*i*) the rash data stream R alone, (*ii*) the rash data stream R divided by the Poisson context data stream P (target/context average $MI = 0.01$), (*iii*) R divided by the context series $P + R$ (target/context average $MI = 0.06$), (*iv*) R divided by the context $P + 2R$ (target/context average $MI = 0.18$), (*v*) R divided by $P + 3R$ (target/context average $MI = 0.32$), and (*vi*) R divided by $P + 4R$ (target/context average $MI = 0.44$).

The results are summarized by the receiver operating characteristic (ROC) curve shown in Fig. 4. ROC curves are plots of the probability of detection (PD) (the fraction of true positives, an empirical sensitivity estimate) versus the probability of false alarm (PFA) (estimated as the fraction of threshold exceedences among the unspiked background data) as the discrimination threshold is varied.[24] As the context time series C included more of the original signal R (from $C = P$ to $C = P + 4R$), the average mutual information between the target P and the context changed from 0.01 to 0.44. From the plotted ROC curves, detection probabilities increased almost uniformly with the target/context MI.

The improvement in signal-to-noise ratio achieved by replacing count data with target/context pairs with positive MI can be expressed directly by the quantity $A_D = PD(T + C) - PD(T)$ at a given PFA value and threshold. The quantity $PD(T)$ is the probability of detection for the target series T alone, and $PD(T + C)$ is the corresponding detection probability for the data stream using the target and context in the ratio T/C. The quantity A_D will be called the "detection advantage," because it summarizes the improvement in alert detection by using a proportion with the target/context pair (T, C) over using the target T alone.

Figure 5 shows the detection advantage A_D for the series of Fig. 4 at $PFA = 0.03$ for both the cross-correlation and mutual information measures. The cross-correlation ρ_{xy} between successive pairs of datasets X and Y was calculated from the expression

$$\rho_{xy} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}, \qquad (5)$$

where E is the expected value operator, and σ_x^2 and σ_y^2 are the respective variances of X and Y. Clearly, the

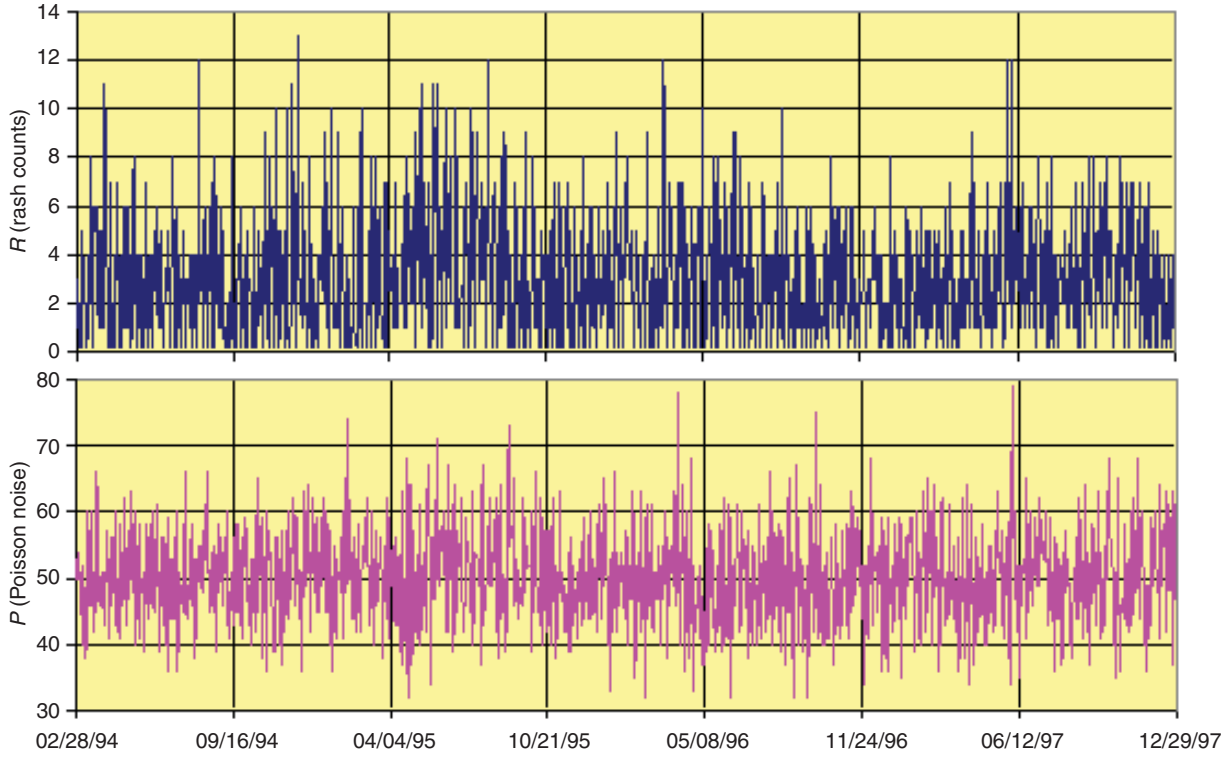

Figure 3. Historical rash counts *R* from chief-complaint syndrome data (upper) used as the target stream to a context data stream *P* of Poisson noise shown with a mean of 50 (lower).

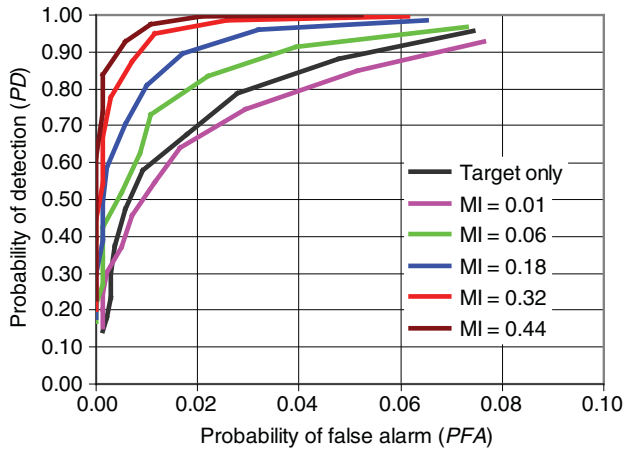

Figure 4. ROC curve constructed from the results of Monte Carlo simulations for a target stream of daily rash syndrome counts when ratios were formed by using context series having increasing mutual information with the target data stream.

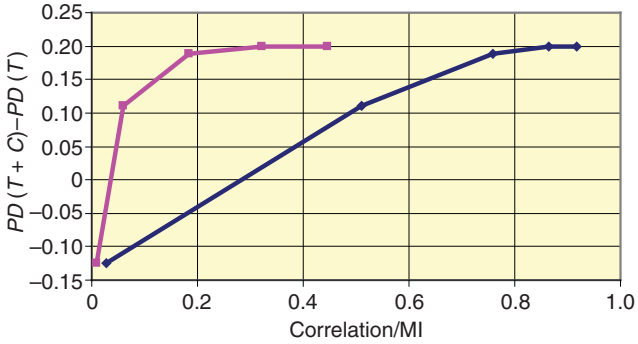

Figure 5. Plots of the detection advantage $AD = PD(T + C) - P(T)$ as a function of average mutual information (magenta) and correlation (blue) for the ROC curves shown in Fig. 4. The detection probabilities were calculated at $PFA = 0.03$. The relative detection probability increases dramatically as correlation and mutual information are increased.

detection probability is enhanced dramatically as the mutual information and correlation between the target and context series increase. For example, the use of the Poisson noise time series (having an *MI* of only 0.01 with the rash count series) as a context gives lower sensitivity than monitoring the rash counts alone. As the mutual information of the context data stream with the target increases, the detection advantage A_D increases sharply to 0.2, giving a 20% increase in the probability of detection.

Additional sets of simulations were conducted with 35 other syndromic count series from the same dataset of outpatient clinic visits. In Fig. 6, we show a set of ROC curves using counts of a botulism-like syndrome as the target and 10 other syndromic series (including the sum of all syndromic visits) as contexts. The black curve is the ROC for the "target-only" series of undivided counts.

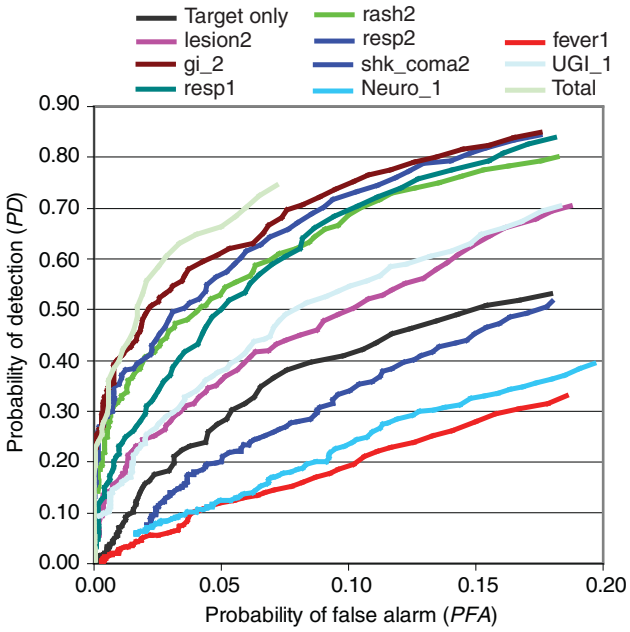

Figure 6. ROC curve obtained by varying the detection threshold over a set of Monte Carlo simulations when the syndrome "Bot-Like-2" is used as a target with 10 other health care data streams.

Note that the three curves below this target-only curve correspond to algorithms applied to ratios using the context data series for neurological, shock/coma, and fever syndrome counts. These three curves indicate lower detection probabilities than the black curve generated by the algorithm applied to counts alone because of their low mutual information ($MI \leq 0.35$) relative to the target count series.

The results of our Monte Carlo simulations on outpatient clinic data can be summarized as follows: (*i*) using *MI* as a metric was slightly better than using correlation, (*ii*) using a shorter time window (e.g., 100 or 200 days) for calculating the mutual information between two time series is preferable to comparing their average *MI* over their entire time history, and (*iii*) the maximum probability of detection among context series occurred almost always when the total syndromic visit count series was used as the context, even if another syndrome data stream had higher *MI* with the target series. This observation is related to coherent integration effects[25] arising from the summation of a large number of (noisy) syndromic data series. This aggregation of many sources causes the average noise amplitude in the total diagnostic series to be reduced by a factor of $\sqrt{N}$, where N is the number of syndromes in the total dataset. This noise reduction is especially evident for highly multivariate datasets like that used in our simulations ($N = 35$).

Practical Implementation of Mutual Information Discrimination

Our *MI* criterion is intended to provide the epidemiologist with a mathematical tool to decide whether to monitor individual counts or to use proportions, and what kind of proportions, when monitoring a public health network.

The Monte Carlo simulations discussed in the previous section suggest that it is useful to use ratios instead of counts to clarify an outbreak if the numerator and denominator of the ratio have sufficient mutual information. Based on results of simulations using an outpatient clinic visit dataset of 35 syndromic time series, our simulations suggest a minimum *MI* of 0.5 for monitoring target/context ratios instead of counts if the method of Daub et al.[10] is used for *MI* calculation. For large datasets, it can be advantageous to use the total diagnostic counts as the context because the noise levels of this data stream are considerably reduced by coherent integration. The advantage of monitoring ratios instead of counts also may be improved by transforming the data series, using short-history time windows, or implementing more sophisticated detection algorithms. Finally, it also is beneficial to use this method over short time windows (e.g., 200 days) to ensure that the measured dependency between target and context series is recent and is not washed out by multiyear averaging.

SIGNAL PROCESSING RLS FILTERS ADAPTED FOR BACKGROUND PREDICTION

Linear Filters for Adaptive Modeling

For useful anomaly detection in biosurveillance, it is necessary to compare recent observed data to an estimate (or predicted value) of what the data ought to be in the absence of a disease outbreak. Therefore, accurate prediction of the data background is an important subtask of automated surveillance. The prediction of syndromic data is complicated by strong nonstationarity and by multiple issues related to data acquisition. Linear filters are particularly useful in the background prediction problem because they are easy to implement and because their adaptive formulation can deal effectively with nonstationary data.[26] The ability to use them in multistream environments is an additional advantage.

A transient event of short duration superimposed on a stationary (or quasi-stationary) background cannot be usefully predicted by using linear prediction filters.[27] However, when recent values of a single data stream are used to predict future values, these predictions may be used as a set of "background" values with respect to which a threshold detector (for unusually high counts) could operate.[27] Accurate background estimates may be expected to yield improved sensitivity at practical alert rates as measured by ROC curves, as shown in the previous section.

Filter Modifications for Biosurveillance Applications

Linear prediction filters are generally easy to implement when the optimality criterion is that of a

minimum least-squared error. We have successfully pioneered and used the adaptive RLS linear filters in the analysis of a "predict and detect" system using over-the-counter medications that were chosen based on their high correlations with a single clinical dataset.[27] The current effort shows that the same RLS prediction method in a univariate (single data stream) environment predicts accurate background counts for a large class of surveillance time series.

A major problem for modeling syndromic data streams from many health indicator information sources is the day-of-week effect. When present, this effect results in a consistent statistical disparity among observed counts on different weekdays. A systematic and adaptive approach to this problem is a vital ingredient in our RLS implementation. Our method to equalize the data distributions relies on an invertible transformation of the data values. The following section describes the day-of-week issue and our solution and implementation. We then show the RLS implementation of the prediction problem and present results and discussion.

Algorithmic Treatment of Weekly Data Patterns

To illustrate the issue, we present military outpatient clinic data from a large metropolitan area. The time series in Fig. 7 represents nearly 4 years of typical daily counts of visits whose diagnoses were classified in the respiratory syndrome. Figure 8 shows the sorted (by value) data corresponding to different days of the week. Weekend (Saturday and Sunday) values fall on curves that are distinctly separate (and lower) from those of the weekdays. Our adaptive solution to this problem is as follows. We fit the middle part of each curve with a straight line and form a reference line using the median of the weekday lines. Other choices for the reference line could work as well.

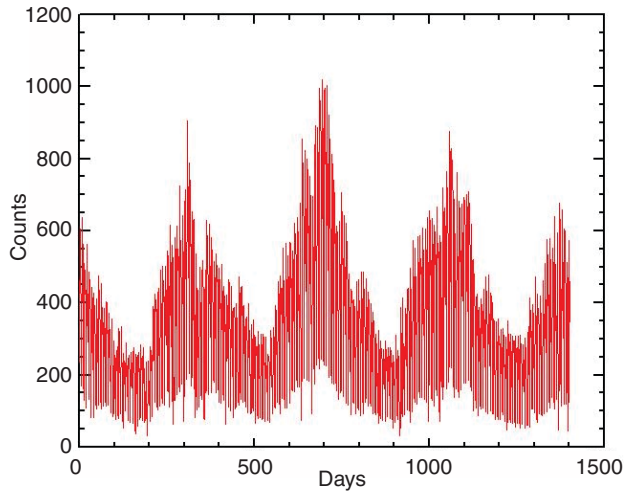

Figure 7. Daily counts of syndromic visits for the respiratory syndrome of military outpatient visits from a large metropolitan area.

All data values then are corrected to this reference line by linear transformation. Figure 9 shows the corrected (sorted) data values after this transformation, and the corrected daily values are shown in Fig. 10. Predictions then proceed on these corrected values. The transformations are inverted by using the same straight line parameters to restore count levels on the original scale.

Background Prediction and the RLS Algorithm

We consider the raw time series of clinical visit counts as both the primary and the reference data channel in order to predict future counts in the following manner. Today's and several past days' counts are combined to make a future clinical data prediction, which then is compared to the actual value of that day's clinical data when that value becomes available, and the error is used

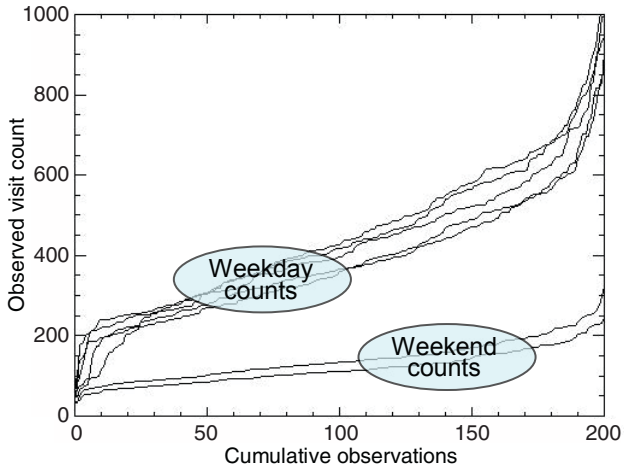

Figure 8. Cumulative frequency distribution (transposed) of daily counts by day of week. The abscissa values are used as indices for day-of-week equalization.

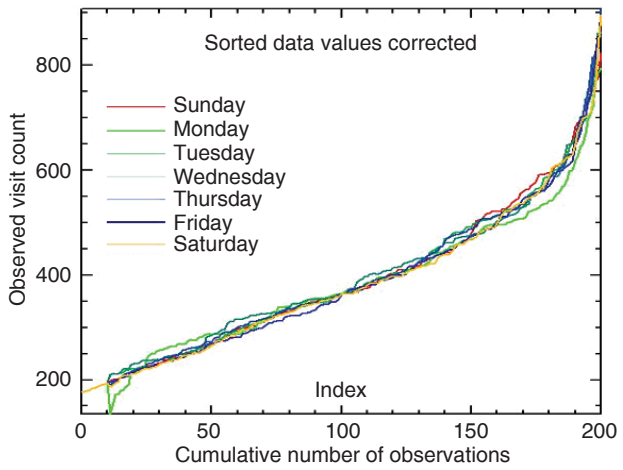

Figure 9. Cumulative frequency distribution (transposed) of daily counts by day of week after equalization.

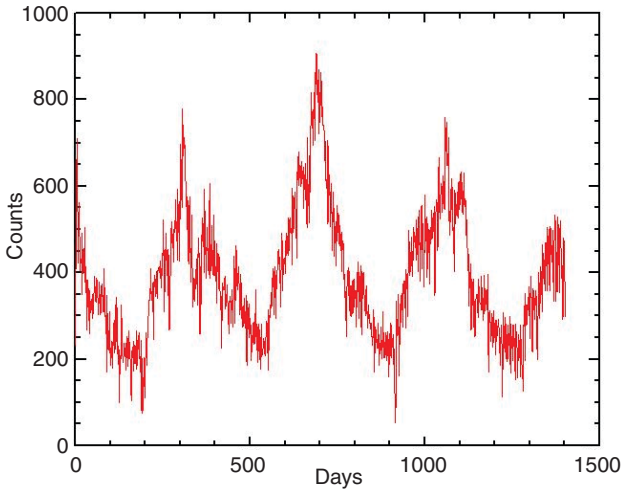

Figure 10. Daily counts of syndromic visits after equalization transformation.

to update the filter coefficients in such a way as to minimize the square of the error.

Denoting the daily counts by y_n, a linear prediction, P days ahead, is given by

$$\hat{y}_{n+P} = \sum_{m=0}^{M-1} h_m y_{n-m} , \qquad (6)$$

where we have assumed a set of linear filter coefficients $\underline{h} = [h_0, h_1, \ldots, h_{M-1}]^T$ of length M. In vector notation, this equation is

$$\hat{y}_{n+P} = \underline{h}^T \underline{y}_n , \qquad (7)$$

where $\underline{h} = [h_0, h_1, \ldots, h_{M-1}]^T$, $\underline{y}_n = [y_n, y_{n-1}, \ldots, y_{n-(M-1)}]^T$, and T denotes matrix transposition. The prediction error is given by

$$e_k = y_k - \hat{y}_k , \qquad (8)$$

and the performance index at day n is

$$\varepsilon_n = \sum_{k=0}^{n} \lambda^{n-k} |e_k|^2 .$$

The "forgetting factor" λ is introduced to deal with nonstationary behavior so that more recent data are given more emphasis.[27] The relationship between the forgetting factor and the effective memory of the filter is found from

$$n_\lambda = \frac{\dot{\sum_{n=0}} n\lambda^n}{\dot{\sum_{n=0}} \lambda^n} = \frac{\lambda}{1-\lambda} .$$

This equation can be solved to give the forgetting factor in terms of the effective memory of the filter, $\lambda = n_\lambda / 1 + n\lambda$, which can be chosen according to observations made on the underlying dataset. We have generally found an effective memory value of 4 weeks (28 days) to give a useful value $\lambda = 0.9655$. The least-squares analogue of the ordinary Wiener filter equations then are

$$\sum_{k=0}^{n} \lambda^{n-k} y_{k-m} y_{k-l} h_l^{(n)} = \sum_{k=0}^{n} \lambda^{n-k} x_k y_m , \qquad (9)$$

where we have introduced the variable $x_n \equiv \hat{y}_{n+P}$. Details of the RLS adaptive algorithm are provided in Box 1.

Note that our implementation uses multiple predictions for the same day, as described above and shown in Fig. 11, in which the data between the left-hand arrows are used to predict each data point (red circle). Thus, given today's and the past available data—i.e., indices $n - (M - 1), \ldots, n$—we make a prediction for day index $n + P$, using the present filters. In addition, we use the same filters to make a prediction for day $n + P - 1$, using the available data indices $[n - 1 - (M - 1), \ldots, n - 1]$, and similarly we continue to make a prediction for every prior day up to and including the index $n + 1$ {the corresponding data indices for the latter prediction are $[n - (M - 1) - (P - 1), \ldots, n - (P - 1)]$}. In other words, every day in the future will have P predictions that were made when that point was P days ahead of the index n, $P - 1$ days ahead, and so on, until it was only 1 day ahead. So when the index n (i.e., today) turns to $n + 1$, i.e., the data for tomorrow become available, we have P possible error terms, only one of which can be fed back into the recursive update equations. In the present implementation, we have found that choosing the error term with the smallest magnitude produces the most effective coefficient updating in terms of prediction accuracy. We experimented with feeding the most recent estimate as well as the one with the smallest magnitude and found the latter to provide slightly better results. In many cases, of course, the most recent estimate also was the one with the smallest magnitude.

Comparative Performance of Adaptive RLS Filter on Syndromic Data

Table 1 lists the daily syndromic time series used to test this method. We used a training period of 1 year to compute reliable estimates of the linear transformation coefficients required in the day-of-week equalization procedure prior to prediction. The last two columns of Table 1 show the median fractional absolute difference for RLS predictions of the listed nine syndromes compared with predictions obtained with a method

BOX 1

The quantities

$$R_{ml}^{(n)} \equiv \sum_{k=0}^{n} \lambda^{n-k} y_{k-m} y_{k-1}$$

and

$$r_m^{(n)} \equiv \sum_{k=0}^{n} \lambda^{n-k} x_k y_m ,$$

which play the roles of the correlation matrix and the cross-correlation vector, satisfy rank 1 update equations

$$\underline{\underline{R}}^{(n)} = \lambda \underline{\underline{R}}^{(n-1)} + \underline{y}^{(n)} \underline{y}^{(n)T}$$
$$\underline{r}^{(n)} = \lambda \underline{r}^{(n-1)} + x_n \underline{y}^{(n)} ,$$

where the superscript n denotes the present day number (with 0 indicating the first day). The least-squares solution to a given a pair $\{\underline{\underline{R}}^{(n)}, \underline{r}^{(n)}\}$ is clearly

$$\underline{h}^{(n)} = [\underline{\underline{R}}^{(n)}]^{-1} . \underline{r}^{(n)} ,$$

and the corresponding prediction value (filter output) is

$$\hat{x}^{(n)} = \underline{h}^{(n)T} . \underline{y}^{(n)} .$$

The RLS algorithm[26] is an adaptive method to relate the solution to the least-squares problem at step n to the solution of the least-squares problem at step $n + 1$. In the parlance of Kalman filtering theory (sequential estimation theory), the quantities at step n (present) are referred to as the *a priori* values, whereas those at step $n + 1$ (next step) are *a posteriori* quantities. An important step in the RLS algorithm is the computation of the inverse correlation matrix, which by using the matrix inversion lemma can be written as follows:

$$[\underline{\underline{R}}^{(n+1)}]^{-1} = [\underline{\underline{R}}^{(n)}]^{-1} - \mu^{(n)} \underline{K}^{(n)} \underline{K}^{(n)T}$$

in terms of the Kalman gain vectors

$$\underline{K}^{(j)} \equiv [\underline{\underline{R}}^{(j)}]^{-1} \underline{y}^{(n)}, \; j = n, n+1$$

and the likelihood variable $\mu^{(n)} = (1 + \underline{y}^{(n)T} . \underline{K}^{(n)})^{-1}$.

The filter update equations are obtained by applying $\underline{\underline{R}}^{(n+1)}$ (i.e., the correlation function of the least-squares problem at step $n + 1$) to the left of $\underline{h}^{(n)}$ (i.e., the solution to the least-squares problem at step n):

$$\underline{h}^{(n)} = \underline{h}^{(n-1)} + e^{(n|n-1)} \mu^{(n)} \underline{K}^{(n|n-1)}$$

while the prediction errors are related to each other by

$$e^{(n)} = \mu_n e^{(n|n-1)} .$$

The hybrid quantities are defined as follows:

$$\hat{x}^{(n|n-1)} = \underline{h}^{(n-1)T} . \underline{y}^{(n)}$$
$$e^{(n|n-1)} = x_n - \hat{x}^{(n|n-1)}$$
$$\underline{K}^{(n|n-1)} = \lambda^{-1} [\underline{\underline{R}}^{(n-1)}]^{-1} . \underline{y}^{(n)}$$
$$\mu^{(n)} = (1 + \underline{K}^{(n|n-1)T} . \underline{y}^{(n)})^{-1} = 1 - \underline{K}^{(n)T} . \underline{y}^{(n)} .$$

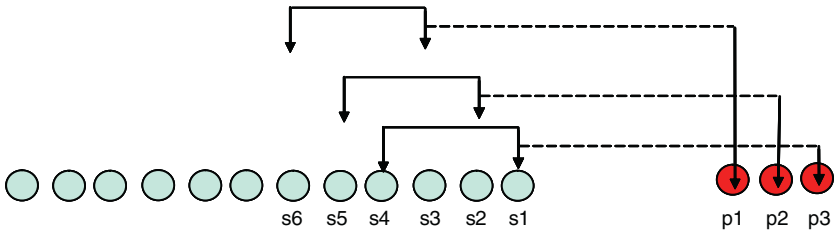

Figure 11. Schematic illustration of prediction algorithm (e.g., values on days s1–s4 are used to predict p3).

Table 1. Median fractional absolute forecast error for nine syndromic time series comparing the CDC W2 algorithm to adaptive RLS.

Syndrome	Mean count	CDC W2 (%)	RLS (%)
Respiratory 1	335	13	9
Respiratory 2	162	16	10
Fever 1	79	16	14
Rash 2	78	18	17
Gastrointestinal (GI) 2	62	16	15
Lower GI 2	55	22	16
GI 1	53	16	15
Lower GI 1	38	18	15
Neurological 2	35	19	18

currently used in the CDC BioSense system[28] for monitoring public health on the national level. This reference method is an enhancement of the Early Aberration Reporting System (EARS) C2 algorithm,[29] denoted W2, in which separate C2 implementations are applied to normal weekdays and to weekend/holidays. For the syndromic visit count series used, note that the RLS fractional prediction errors are consistently below those of the reference W2 method. Importantly, for day-to-day monitoring experience, this advantage holds uniformly when comparisons are stratified by day of week and by month of year.

The modified RLS filter-based predictor presented above yielded useful predictions for a variety of city-level syndromic data series. More recent comparisons suggest

that it may be useful on smaller scales as well, and efforts are ongoing to establish the class of time series for which it is useful and to select this predictor based on a modest set of historic data. A logical next step is to investigate the advantage in signal detection capability that may be obtained by using residuals calculated with this specialized RLS adaptation. Both simple threshold detectors and more complex control charts are to be applied for this purpose.

PRACTICAL ADAPTATION OF TEMPORAL ALERTING METHODS FOR ROUTINE OPERATIONS

Basic challenges and common approaches used in univariate temporal alerting algorithms for biosurveillance have been discussed previously in this journal.[1] This section discusses advances and adaptations that have resulted from experience working with epidemiologists, from the changing data environment, and from recent research.

Adaptive Choice of Detection Algorithms

The use of data models and control charts for disease surveillance was discussed previously in this journal.[1] Although the adaptive regression approach discussed in that article has proved applicable to many facility-level and county-level data streams, this modeling has little explanatory value for sparser time series that occur when data records are filtered more finely because of geographic restrictions, concerns for particular age groups, or subsyndrome classifications. Moreover, ESSENCE has evolved to allow users to choose ad hoc combinations of data filters, and rapid anomaly detection of the resulting time series is required. The first and currently implemented solution to the dilemma of whether and how to model was an automated choice between adaptive regression and an exponentially weighted moving average (EWMA) chart, discussed next. A more recent, unified solution uses generalized exponential smoothing, discussed in the section on Holt–Winters-based control charts.

The current automated algorithm selection is determined by a goodness-of-fit test to determine whether the adaptive regression model has explanatory value. For each day's baseline, regression coefficients are refit by using standard predictor variables. For a goodness-of-fit measure, the adjusted R^2 coefficient estimates the portion of the sum of squared baseline modeling errors that are explained by the predictor variables. If this measure exceeds 0.6, the system uses the regression model to make a current-day forecast and decide whether the day's observed count is anomalous. Time series of record counts using the more inclusive respiratory and gastrointestinal syndromes usually pass this test in medium to large regions where the median daily count is well above 10. However, if the adjusted R^2 does not exceed 0.6, an adaptive EWMA control chart is applied to test for anomaly.

The experience of recent years of ESSENCE use has led to modifications in both the regression model and EWMA algorithms implementations, and we summarize these modifications below.

Modifications to Adaptive Regression Algorithm

- *Predictor variables*: The various data types used in ESSENCE show several different day-of-week patterns, with many data sources showing characteristic weekend drop-offs ranging from 75% to 100%. Weekly patterns in hospital emergency room data vary widely according to the type of hospital and level of weekend staffing. To accommodate the resulting range of time series behaviors, six day-of-week indicator variables now are included in the model. The other predictor variables are a linear trend term and indicators for holiday and post-holiday effects. By contrast with some research efforts,[30,31] long-term predictors such as harmonic seasonal terms are not included in ESSENCE because the data history often is not sufficient for the use of these terms, and even when it is sufficient, year-to-year changes in information systems, diagnosis coding, population behavior, and even weather can degrade their predictive value.
- *Outlier removal*: Alerting problems have occurred because of anomalous baseline data values even when the goodness-of-fit test selects the regression model. These outlier values may reflect true health events but often result from issues in the data chain, and, regardless of their cause, they should not be used to infer regression coefficients for forecasting. To avoid training on these outliers, observations outside the alerting confidence bounds are replaced by those bounds for model-fitting.
- *Probability-scale representation of algorithm output*: The unscaled outputs of the regression algorithm are the forecast errors divided by the standard error of regression. For alerting purposes, values representing degree of anomaly are derived from Student's t test distribution lookups using these outputs with the number of degrees of freedom set to the baseline length minus the number of predictor variables. Note that the resulting p values are not outbreak probabilities but only statistical anomaly measures on a 0–1 scale. Use of this scale allows comparison of outputs among algorithms and direct combination of these outputs by multiple univariate methods.

Modifications to the EWMA Control Chart Algorithm

When the regression goodness-of-fit test fails in ESSENCE data analysis, an adaptive version of the

EWMA control chart of statistical process control (SPC) is applied.[32] The SPC formulation assumes that the input data stream X_t is Gaussian with mean μ and variance σ^2. The ESSENCE test statistic is $(Z_t - \bar{x}_t)/ks_t$, where Z_t is the current weighted moving average

$$Z_t = \omega X_t + (1 - \omega)Z_{t-1} , \tag{10}$$

and ω is a smoothing constant between 0 and 1. The sliding baseline mean $\bar{x}_t$ and standard deviation s_t give estimates of μ and σ. The denominator ks_t approximates the standard deviation of Z_t, where the constant k is given by

$$k^2 = \left(\frac{\omega}{(2-\omega)}\right)(1-(1-\omega)^{2t}) . \tag{11}$$

The ESSENCE implementation uses a 28-day sliding baseline and a 2-day guard band, or buffer, separating the baseline from the current day.

Adaptations to EWMA Chart Implementation

The following modifications have been adopted to improve the accuracy and robustness of statistical alerting in response to concerns of epidemiologist users, typical data behavior, and the most common problems in the data acquisition chain.

(i) Sensitivity to both sudden and gradual signals: Data signals expected from infectious disease outbreaks are not the lasting step increases that one would expect of a mean shift. The signals of interest are transient data effects of epidemic curves of attributable cases lasting from a few days to a month or more. Even for a given disease such as influenza, the outbreak signal may be sudden and explosive or more gradual. In a published hospital-based application, cumulative sum (CUSUM)-Shewhart and EWMA-Shewhart charts were applied to detect both types of signal in the monitoring of hospital infections.[33] Those authors found an EWMA-Shewhart chart preferable for autocorrelated background data. Emulating this strategy, the ESSENCE EWMA algorithm is applied for smoothing coefficients of both 0.4 and 0.9, with the smaller coefficient for sensitivity to gradual signals and the larger one to approximate a Shewhart chart for sensitivity to spikes.

(ii) Correction for the adaptive baseline: The conventional EWMA statistic was developed for stationary Gaussian data, and the use of the sliding window for daily parameter adjustment changes the variance of the weighted moving average from the simple, fixed-parameter situation. From taking the variance of $(Z_t - \bar{x}_t)$ and expanding, the total adjusted variance at time step j is the s_t^2 times the factor

$$\left(\frac{\omega}{(2-\omega)}\right)(1-(1-\omega)^{2j}) + (1/B) - 2(1-\omega)^{g+1}\left(\frac{(1-\omega)^B - 1}{B}\right) \tag{12}$$

for baseline length B and guard band length g. Therefore, the sample standard deviation s_t is multiplied by the square root of this factor in the test statistic.

(iii) Probability-scale representation of algorithm output: As in the regression model implementation, this modification was partly a cultural one to bridge the gap between classical epidemiologist practice and SPC control chart usage. Given the modification *ii* for the adaptive baseline, probability values representing the degree of anomaly are derived from Student's t test distribution lookups with the number of degrees of freedom set to the baseline length – 1, assuming that the input data are Gaussian. Again, the resulting p values should not be interpreted as outbreak probabilities.

(iv) Bounding the baseline variance: This modification was needed because of the many varied time series monitored for both general syndromic and more diagnosis-specific surveillance. Because of the multiple strata used as data filters, many of the monitored time series are sparse, with median daily counts of zero. Thus, for a 4- to 8-week sliding baseline, the sample standard deviation s_t of these series often is near zero, and for detection statistics analogous to $(x - u)/s_t$, near-zero variances cause instability and unwanted alerts. In general, single isolated cases are not signals. Therefore, health monitors were asked which small temporal clusters should cause alerts and which should not. Minimum variances were established from this guidance by algebraic manipulation of the EWMA test statistic.

(v) Small-count corrections for monitoring non-Gaussian time series: Despite the adjustments above, for many time series the algorithms were producing too many alerts because series values were not Gaussian-distributed. In biostatistics applications, count data often are assumed to obey a Poisson distribution, for which the variance equals the mean value. The distribution of syndromic count data is usually closer to Poisson than to Gaussian in that the dependence of the variance on the mean is evident (though overdispersion evidenced by exaggerated variances often is seen because the counts are not from homogeneous populations). For this reason, false alarms were reduced by adding a factor c/s_t to each Gaussian-derived alerting threshold, i.e., by adding a fixed, empirically derived constant c divided by the data standard deviation. This approach produces adjustments that are significant for small-count time series but negligible for series with larger counts. For Poisson data streams, this approach yielded expected probabilities of threshold exceedence for mean count

values at least as small as 0.1 per day. The threshold adjustment is not directly applicable to the EWMA strategy because the EWMA Z_t of a Poisson variable is not still Poisson. However, the correction applied to Z_t/ω again produced the expected probabilities for smoothing constants $\omega = 0.4$ and 0.9 when the correction constant c was empirically computed as a function of both ω and the desired p value α. Optimizing this function over a range of mean values from 0.1 to 20 gave the following formula for c:

$$c(\omega, \alpha) = 0.1304 - (0.2409 - 0.1804 \times (1 - \omega)^4) \times \log(10 \times \alpha) \, . \quad (13)$$

Figure 12 illustrates the effect of this correction.

Combining *ii* through *v*, the corrected adaptive EWMA alerting method for smoothing constants $\omega = 0.4$ and 0.9 is

- At step j, compute the test statistic $Z_t = (E_t - \bar{x}_t)/\max(s_t^*, s_{\min})$, where E_t is the current EWMA, $\bar{x}_t$ is the current baseline mean, s_t^* is the baseline standard deviation adjusted as in *ii*, and $s_{\min}$ is the empirical minimum in *iv*.
- Form the adjusted test statistic Z_t^* by subtracting the non-Gaussian correction term

$$Z_t^* = Z_t - c(\omega, \alpha) \times \omega/(\omega, \alpha) \times \omega/\max(s_t^*, s_{\min}) \, . \quad (14)$$

- Alert if Z_t^* exceeds the critical t distribution value for $B - 1$ degrees of freedom at p value α, where B is the baseline length.

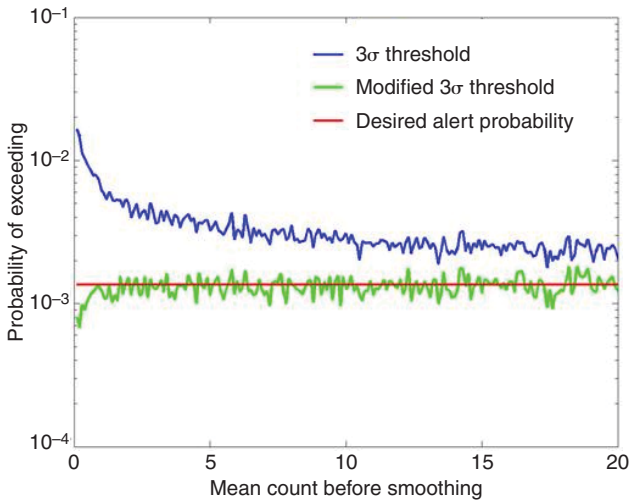

Figure 12. Monte Carlo estimates of false-alert probabilities for an EWMA parameter $\omega = 0.4$. The red line is desired alert probability, the blue curve is the result if the EWMA input data are Poisson-distributed, and the green curve is the result for the same inputs if the threshold is adjusted to $3 + 1.07\,\omega/s$.

(vi) Automated management of apparent data drop-outs: Some algorithm problems reported by ESSENCE users have resulted from data acquisition or reporting problems, and these were frequently caused by data drop-outs. Typically, a clinic's information system goes down for a week or more, and data from the affected facilities are interrupted until transmission to ESSENCE is restored. Sometimes corrected previous-day values become available, but often the system retains zero values for those days. Statistical detection and adjustment for these problems is difficult because full and partial drop-outs may cause a variety of "errors" in the sense of irrelevant alarms. The simplest resulting problem is that the sliding baseline mean and standard deviation assume unrealistic values during the drop-out, and meaningless alerts are seen when good data start up again. This situation arises from a common informatics limitation that missing data often are impossible to distinguish from zero values because of the large number of data suppliers and confidentiality restrictions in biosurveillance systems. Therefore, the following statistical solution has been adopted. Suppose that the baseline contains a string of M zeros and that Eq. 10 is the set of values outside this string. The M zeros are treated as missing data, and the baseline is restarted if

$$(NZ(B1)/\text{length}(B1))^M < \alpha \, , \quad (15)$$

where $NZ(B1)$ is the number of zeros in Eq. 10 and α is a threshold probability, with $\alpha = 0.01$ adopted from experience. Thus, the string of zeros is more likely to be treated as missing data if few zeros appear in the rest of the counts. The implemented logic is more complex to account for multiple zero strings, but this basic idea avoids nuisance alarms and quickens algorithm recovery time.

Limitations of Automated Regression/Control-Chart Approach

The simple algorithm selection criterion combined with improvements to the regression and EWMA algorithms has helped reduce data/algorithm mismatch and false-alarm rates in the environment of increasingly complex data and surveillance objectives. However, limitations of this approach cause occasional anecdotal problems, and the sheer number of data streams to monitor calls for continued improvements and system-level algorithm combinations.

The most frequent problems result from violations of the underlying data assumptions. For example, regression residuals in syndromic time series often are autocorrelated, variances are nonhomogeneous, and means are not constant as a function of time. These realistic data issues can cause the regression and EWMA methods to scale anomalies differently. More advanced modeling

approaches like weighted and robust regression, autoregressive integrated moving average (ARIMA) modeling, and transformation of the dependent variables can reduce these problems,[31,34] but often they introduce more assumptions and more coefficients to be estimated. It is usually suggested to have at least five observations for robust estimation of every parameter in regression model. Thus, predictors in the current ESSENCE model suggest at least 8 weeks of data in the baseline, and much longer characteristic data baselines are available from relatively few data sources. Furthermore, the advantages of more complex modeling may come at the cost of resource-intensive data analysis,[35] and such analysis often is impractical for numerous, ad hoc data streams.

Anecdotal problems have been caused by artifacts of the fixed-length moving baseline and the exact goodness-of-fit criterion for choosing regression modeling. One approach that is more flexible than regression but can accommodate seasonality, short-term trends, and data peculiarities is generalized exponential smoothing, which is widely used for financial forecasting and introduced in the next section.

Data Forecasting Using Generalized Exponential Smoothing

The idea of exponential smoothing has been extended to the modeling of local changes in trend[36] and then in periodic behavior.[37] In the well-known Holt–Winters implementation, updating equations analogous to Eq. 10 are combined to obtain forecasts accounting for changes in process mean, seasonal components, and trend behavior. Specifically, let α, β, and γ denote smoothing coefficients (where α has the role of ω in simple EWMA) for updating terms corresponding to the level, trend, and seasonality, and let s be the cycle length, in data time steps, of seasonal or periodic behavior in the data series. Forecast components m_t, b_t, and c_t then are updated with the following equations.

Level:
$$m_t = \alpha \frac{y_t}{c_t - s} + (1-\alpha)(m_{t-1} + b_{t-1}), \quad 0 < \alpha < 1 \tag{16}$$

Trend:
$$b_t = \beta(m_t - m_{t-1}) + (1-\beta)b_{t-1}, \quad 0 < \beta < 1 \tag{17}$$

Seasonality:
$$c_t = \gamma \frac{y_t}{m_{t-1^*}} + (1-\gamma)c_{t-s}, \quad 0 < \gamma < 1 \tag{18}$$

Eq. 18 contains a recent, robust improvement with the use of m_{t-1}.[38] In the multiplicative Holt–Winters procedure, these components are combined to yield a k step-ahead forecast:

$$\hat{y}_{n+k|n} = (m_n + kb_n)(c_{n-s+k}). \tag{19}$$

Holt–Winters Forecasting for Biosurveillance

Forecasting by the Holt–Winters method has been applied[39] to the prediction of daily biosurveillance data using 16 authentic data streams on the scale of Fig. 7. In this study, the cycle length s was set at 7 to account for the frequent but data-source-dependent weekly patterns reported above for daily syndromic series, and Holt–Winters forecasts gave forecast errors that consistently were lower than those obtained with an adaptive regression model. We have applied Holt–Winters forecasts to build an anomaly detector that is robust with respect to data scale and to common data features such as holiday effects. We summarize the modifications required for robust forecasting and then present the adaptive Holt–Winters control chart.

(i) Updating adjustments for predictable outliers: Updating is applied with the usual cyclic multiplier replaced by a special factor for holidays or other calendar events.
(ii) Updating adjustments for unexpected outliers: Updating of the component terms is suspended if the current data are anomalous according to an absolute fractional error criterion. This condition has been refined across data scales to avoid spurious training.
(iii) Adjustment for sparse time series: Data with sequential zeros may cause an imbalance in component updating that leads to artificial trending. To avoid such artifacts, a small data-dependent constant is added to all input values and then subtracted from the composite forecast.
(iv) Choice of initial values for level, trend, and cyclic components: Initial value effects are negligible in simple EWMA but are known to be critical in more generalized Holt–Winters forecasting.[35] Experience with syndromic series forecasting has confirmed this finding and shown that grossly misspecified initial values may degrade algorithm performance for several months using daily data and also may confound the choice of smoothing coefficients α, β, and γ. From experience with numerous syndromic series on multiple scales, the recommended starting values for level, trend, and seasonal multipliers are the overall baseline mean, an initial zero slope, and stratified averages for the weekly multipliers, respectively. If no historic data are available, initialization should be based on the best information or

experience at hand. After 4–6 weeks of data have been collected, early guesses for the initial values should be replaced with values for the recommended computations.

(v) Choice of smoothing coefficients: The selection of Holt–Winters smoothing coefficients is critical for reliable forecasting. Based on forecast results using a variety of data streams and data-grouping criteria, we categorized input series by median value using categories of *sparse* (median < 1), *low* (1 < median < 10), *average* (10 < median < 100) , or *high* (median > 100). Following previous studies,[35,39] we applied a coarse, multidimensional grid search to determine an effective set of smoothing coefficients for each median category. Figure 13 illustrates the results of this search for combinations of α and γ, for $\beta = 0$. Each color-coded cell represents the goodness-of-fit-based rank of a particular [α, γ] coefficient pair. Blue represents the best pairs as indicated by the highest rank. The sets of coefficients shown in Table 2 were chosen for each group.

Table 2. Coefficient sets.

Name	Median	Chosen [α, β, γ] coefficient set
Sparse	0	[0.05, 0, 0.10]
Low	[1,10)	[0.05, 0, 0.05]
Average	[10,100)	[0.15, 0, 0.05]
High	[100, ∞)	[0.30, 0, 0.05]

Forming an Alerting Algorithm from the Residuals

Even for reliable data forecasts, the forecast residuals, or differences between observed and expected values, are not sufficient for anomaly detection. Estimates of residual variance also are needed to calculate a detection statistic in order to reduce day-to-day changes in sensitivity and specificity. For example, in syndromic data streams, a day-of-week effect is present not only in sample data means but also more subtly in residual variances, despite forecast efforts to remove this effect. We tested and adopted a method suggested by Chatfield[40] and nearly identically by Koehler[38] in which variance is a function of level and cyclic parameters.[21] The Koehler estimate provided the detection statistic $(y_{t+k} - \hat{y}_t(k))/\sqrt{Var\,(e_t(k))}$ with the closest fit to a standard Gaussian distribution.

Holt–Winters-Based Algorithm Performance

A study has been completed to compare the detection performance of the above statistic to the traditional regression-based detector. Simple 1-day spikes were added to authentic background data streams previously collected in ESSENCE. Probabilities of detecting the injected spikes were tabulated for the new normalized Holt–Winters statistic and for the adaptive regression method, respectively. Average detection probabilities,

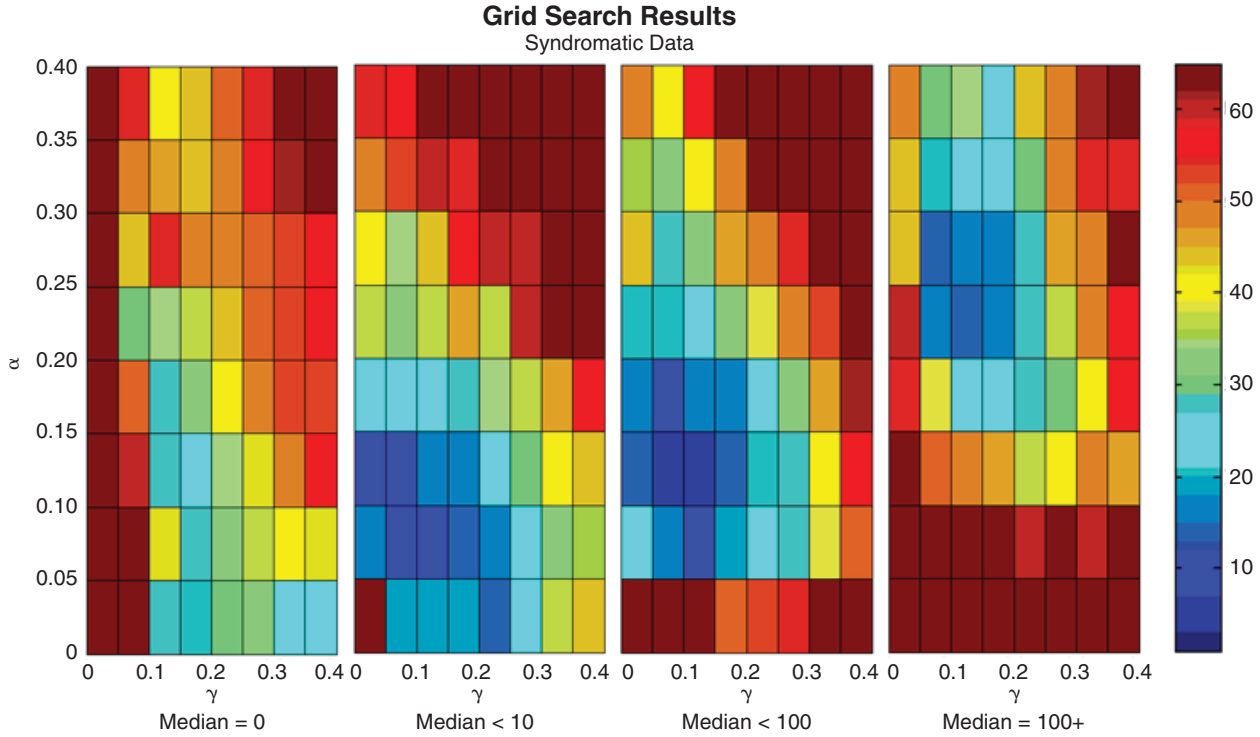

Figure 13. Results of grid searches for optimal sets of smoothing coefficients in each data category.

equivalent to the scaled area under the relevant part of the ROC curve, were calculated for background alert levels ranging from 1 per 28 days to 1 per 56 days. Table 3 summarizes the results of this study and shows a consistent detection performance advantage for the new Holt–Winters method. Although it is probably demonstrable that more sophisticated modeling could improve the regression detector for any given time series, such efforts must keep in mind the variety of data types, limited on-site analysis capability, and other requirements of biosurveillance systems.

With the success of these preliminary results, work is in progress on evaluating Holt–Winters detection performance for detection of signals of varying realistic shape and duration for ESSENCE implementation.

SUMMARY AND RESEARCH DIRECTIONS

This article shows how challenges of developing automated disease surveillance systems are being approached by statistical means. The discussion has been limited to univariate alerting methods, but current ESSENCE methods also include multivariate statistical alerting for simple data fusion[41] and scan statistics for detection of localized case clusters.[1]

A considerable body of research has addressed syndromic classification of clinical patient records in order to clarify potential data signals for early warning of disease outbreaks.[6] Subsequent to these classification decisions are questions of how to form the most useful outcome variables and how to monitor them. Research issues include how to best combine and transform the derived time series. The *MI* criterion presented above gives a means for evaluating target/context quotients, typically a series of daily syndromic facility visit counts divided by the daily facility visit total, for their capacity to increase the signal-to-noise ratio for the syndrome of interest. The *MI* simulation results suggest that appropriately conditioning and combining syndromic data streams can be important in achieving the signal-to-noise gains necessary for improved detection performance in biosurveillance. Further research is continuing on applying optimal filtering to the target/context problem and on implementing shorter time windows for more discriminating alert detection. Information-theoretic methods, little used in disease surveillance up to now, may have additional potential utility, such as application to larger sets of data streams to select advantageous combinations for monitoring by adaptive multivariate SPC.

The requirement for adaptive data background estimation in biosurveillance is driven by the need for increasingly early signal recognition. Before the late 1990s, most disease surveillance was done on weekly, monthly, or longer time scales. More recently, daily monitoring

Table 3. Comparison of empirical detection probabilities obtained from Holt–Winters and adaptive regression methods.

Syndromic data series	Holt–Winters detection probabilities	Adaptive regression detection probabilities	Median daily data count
Resp_1	0.988	0.984	310.5
Resp_2	0.994	0.983	184
Fever_1	0.921	0.879	90
Rash_2	0.994	0.985	75
Gi_2	0.990	0.985	72
Lesion_2	0.979	0.960	62
GI_1	0.989	0.945	55
Lesion_1	0.923	0.943	43
Neuro_2	0.986	0.969	39
LGI_2	0.920	0.870	23
UGI_1	0.910	0.822	15
Hemr_ill_2	0.944	0.863	15
Bot_Like_2	0.957	0.882	15
UGI_2	0.888	0.832	8
Lymph_1	0.828	0.724	7
Shk_Coma_2	0.579	0.517	5
Bot_Like_1	0.645	0.509	3
Rash_1	0.537	0.484	3

Average probabilities are shown for practical background alert rates for a signal level of twice the data standard deviation of each data series.

of population data has become common among health departments, and advances in hospital informatics systems are pushing analysts to develop near real-time alerting capability. We have shown how linear RLS filters may be modified for improved prediction of syndromic time series, and we obtained consistent improvements in prediction for a variety of city-level data series.

Efforts are currently in progress to specify the class of time series for which these filters are most useful and to calculate filter coefficients from limited data history. Another important extension is the use of RLS filters for multivariate background prediction, an effort that is promising because of the potential to adapt quickly to changes in the multistream covariance matrix. Such changes present a major obstacle for other modeling approaches.

The final section introduced anomaly detection methods based on forecasts using generalized exponential smoothing. Engineering modifications were presented to get robust detection performance from Holt–Winters forecasts across a large set of disparate syndromic time series.

We summarize the advantages of the Holt–Winters-based adaptive control charts:

- The normalized Holt–Winters detector outperforms traditional regression-based method on most syndromic data streams.
- As in the application of simple EWMA charts, robust detection performance does not depend on assumptions of normality, stationarity, and constant variance.
- EWMA is a special case of Holt–Winters smoothing, so the specific adjustments described above for sparse data streams can be included in a more general Holt–Winters framework.
- Only a limited amount of data is required to initialize the algorithm, and the entire data stream can be used for detection analysis.
- With the choice of a single, adaptive method for each time series category, the automated switching among algorithms is eliminated, avoiding occasional problems from the anomaly scale or from incorrect switching.
- The residual variance estimation presented reflects natural properties of syndromic data. No significant autocorrelation is left in the forecast residuals.

Following the success of these preliminary results, further testing of adaptive Holt–Winters control charts is underway for detection of signals of varying realistic shape and duration. The varying objectives, data environments, and geographic scales of modern biosurveillance systems pose numerous obstacles that can be solved only by a combination of epidemiological and informatics advances along with analytical ones like those of the current study.

ACKNOWLEDGMENTS: This work was supported by Grant 1-R01-PH-24-1 from the CDC. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the CDC.

REFERENCES

[1]Burkom, H. S., "Development, Adaptation, and Assessment of Alerting Algorithms for Biosurveillance," *Johns Hopkins APL Technical Digest*. **24**(4), 335–342 (2003).
[2]Hurt-Mullen, K. J., and Coberly, J., "Syndromic Surveillance on the Epidemiologist's Desktop: Making Sense of Much Data," *MMWR Morb. Mortal. Wkly. Rep.* **54**(Suppl.), 141–146 (2005).
[3]Burkom, H., "Alerting Algorithms for Biosurveillance," Chap. 6, in *Disease Surveillance: A Public Health Informatics Approach*, J. S. Lombardo and D. L. Buckeridge (eds.), John Wiley & Sons, Inc., Hoboken, NJ, pp. 143–192 (2007).
[4]Reis, B. Y., Kohane, I. S., and Mandl, K. D., "An Epidemiological Network Model for Disease Outbreak Detection," *PLoS Medicine*. **4**(6), e210 (2007).
[5]Gordia, L., *Epidemiology*, Second Ed., W. B. Saunders, Philadelphia, PA (2000).
[6]Babin, S., Magruder, S., Hakre, S., Coberly, J., and Lombardo, J. S., "Understanding the Data: Health Indicators in Disease Surveillance," Chap. 2, in *Disease Surveillance: A Public Health Informatics Approach*, J. S. Lombardo and D. L. Buckeridge (eds.), John Wiley & Sons, Inc., Hoboken, NJ, pp. 43–90 (2007).
[7]Wojcik, R., Hauenstein, L., Sniegoski, C., and Holtry, R., "Obtaining the Data," in *Disease Surveillance: A Public Health Informatics Approach*, J. S. Lombardo and D. L. Buckeridge (eds.), John Wiley & Sons, Inc., Hoboken, NJ, pp. 91–142 (2007).
[8]Cover, T. M., and Thomas, J. A., *Elements of Information Theory*, John Wiley & Sons, Inc., Hoboken, NJ (2006).
[9]Herwig, R., Poustka, A. J., Meuller, C., Lehrach, H., and O'Brien, J., "Large-Scale Clustering of cDNA-Fingerprinting Data," *Genome Res.* **9**, 1093–1105 (1999).
[10]Daub, C. O., Steuer, R., Selbig, J., and Kloska, S., "Estimating Mutual Information Using B-Spline Functions—An Improved Similarity Measure for Analysing Gene Expression Data," *BMC Bioinformatics* **5**, 118 (2004).
[11]Priness, I., Maimon, O., and Irad, B.-G., "Evaluation of Gene-Expression Clustering via Mutual Information Distance Measure," *BMC Bioinformatics* **8**, 111 (2007).
[12]Fraser, A. M., and Swinney, H. L., "Independent Coordinates for Strange Attractors From Mutual Information," *Phys. Rev. A* **33**(2), 1134–1140 (1986).
[13]Sagar, R. P. and Guevara, N. L., "Mutual Information and Electron Correlation in Momentum Space," *J. Chem. Phys.* **124**, 134101 (2006).
[14]Nichols, J. M., Moniz, L., Nichols, J. D., Pecora, L. M., and Cooch, E., "Assessing Spatial Coupling in Complex Population Dynamics Using Mutual Prediction and Continuity Statistics," *Theor. Popul. Biol.* **67**(1), 9–21 (2005).
[15]Kraskov, A., Stogbauer, H., and Grassberger, P., "Estimating Mutual Information," *Phys. Rev. E* **69**(6), 066138 (2004).
[16]Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J., "The Mutual Information: Detecting and Evaluating Dependencies Between Variables," *Bioinformatics*. **18**(Suppl. 2), S231–S240 (2002).
[17]Sturges, H. A., "The Choice of a Class Interval," *J. Am. Stat. Assoc.* **21**, 65–66 (1926).
[18]Law, A. M., and Kelton, W. D., *Simulation Modeling and Analysis*, McGraw–Hill, New York (1997).
[19]Scott, D. W., "On Optimal and Data-Based Histograms," *Biometrika* **66**(3), 605–610 (1979).
[20]Moon, Y.-I., Rajagopalan, B., and Lall, U., "Estimation of Mutual Information Using Kernel Density Estimators," *Phys. Rev. E* **52**(3), 2318–2321 (1995).
[21]Witten, I. H., and Frank, W., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers Inc., San Francisco, CA (2000).
[22]Michaels, G., Carr, D., Askenazi, M., Furman, S., Wen, X., and Somogyi, R., "Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data," *Pac. Symp. Biocomput.* **3**, 42–53 (1998).

[23]Strehl, A., and Ghosh, J., "Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions," *J. Machine Learn. Res.* **3**(3), 583–617 (2002).
[24]McNicol, D., *A Primer of Signal Detection Theory*, George Allen & Unwin, London (1972).
[25]Mahafza, B., *Radar Systems Analysis and Design Using MATLAB*, Chapman & Hall/CRC, Boca Raton, FL (2000).
[26]Moon, T. K., and Stirling, W. C., *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, Upper Saddle River, NJ (2000).
[27]Najmi, A. H., and Magruder, S. F., "An Adaptive Prediction and Detection Algorithm for Multistream Syndromic Surveillance," *BMC Med. Inform. Decis. Mak.* **5**, 33 (2005).
[28]Bradley, C. A., Rolka, H., Walker, D. and Loonsk, J., "BioSense: Implementation of a National Early Event Detection and Situational Awareness System," *MMWR Morb. Mortal. Wkly. Rep.* **54**(Suppl.), 11–19 (2005).
[29]Hutwagner, L., Thompson, W., Seeman, G. M., and Treadwell, T., "The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS)," *J. Urban Health* **80**, i89–i96 (2003).
[30]Brillman, J. C., Burr, T., Forslund, D., Joyce, E., Picard, R., and Umland, E., "Modeling Emergency Department Visit Patterns for Infectious Disease Complaints: Results and Application to Disease Surveillance," *BMC Med. Inform. Decis. Mak.* **5**(4), 1–14 (2005).
[31]Craigmile, P. F., Kim, N., Fernandez, S., and Bonsu, B., "Modeling and Detection of Respiratory-Related Outbreak Signatures," *BMC Med. Inform. Decis. Mak.* **7**, 28 (2007).
[32]Ryan, T. P., *Statistical Methods for Quality Improvement*, John Wiley & Sons, Inc., New York (1989).
[33]Morton, A. P., Whitby, M., McLaws, M.-L., Dobson, A., McElwain, S., et al., "The Application of Statistical Process Control Charts to the Detection and Monitoring of Hospital-Acquired Infections," *J. Qual. Clin. Prac.* **21**, 112–117 (2001).
[34]Shmueli, G., and Fienberg, S. E., "Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Biosurveillance, in *Statistical Methods in Counterterrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, A. G. Wilson, G. D. Wilson, and D. H. Olwell (eds.), Springer, New York (2006).
[35]Chatfield, C., and Yar, M., "Hold–Winters Forecasting: Some Practical Issues," *The Statistician* **37**(2), 129–140 (1988).
[36]Holt, C. E., "Forecasting Trends and Seasonals by Exponentially Weighted Averages," Carnegie Institute of Technology, Pittsburgh, PA, Office of Naval Research Memorandum No. 52.
[37]Winters, P. R., "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Sci.* **6**, 324–342 (1960).
[38]Koehler, A. B., Snyder, R. D., and Ord, J. K., "Forecasting Models and Prediction Intervals for the Multiplicative Holt–Winters Method," *Int. J. Forecast.* **17**, 269–286 (2001).
[39]Burkom, H., Murphy, S. P., and Shmueli, G., "Automated Time Series Forecasting for Biosurveillance," *Stat. Med.* **26**(22), 4202–4218 (2007).
[40]Chatfield, C., and Yar, M., "Prediction Intervals for Multiplicative Holt–Winters," *Int. J. Forecast.* **7**, 31–37 (1991).
[41]Magruder, S. F., Lewis, S. H., Najmi, A., and Florio, E., "Progress in Understanding and Using Over-the-Counter Pharmaceuticals for Syndromic Surveillance," *MMWR Morb. Mortal. Wkly. Rep.* **53**(Suppl.), 117–122 (2004).

The Authors

Howard S. Burkom received a B.S. degree from Lehigh University and M.S. and Ph.D. degrees in mathematics from the University of Illinois at Urbana–Champaign. He has 7 years of teaching experience at the university and community college levels. Since 1979, he has worked at APL developing detection algorithms for underwater acoustics, tactical oceanography, and public health surveillance. Dr. Burkom has worked exclusively in the field of biosurveillance since 2000, primarily adapting analytic methods from epidemiology, biostatistics, signal processing, statistical process control, data mining, and other fields of applied science. He is an elected member of the Board of Directors of the International Society for Disease Surveillance. **Yevgeniy Elbert** received his bachelor's degree in mathematics from Kiev University in the Ukraine. He continued his education at University of Maryland Baltimore County, receiving a master's degree in mathematical statistics in 1999. He worked as a statistician at Walter Reed Army Institute of Research from 2001 before joining APL in 2006. The majority of Mr. Elbert's work is devoted to the development of the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) as part of the APL research team. **Steven F. Magruder** is a member of the APL Principal Professional Staff. He received a Ph.D. in physics in 1978 from the University of Illinois at Urbana–Champaign. He joined APL that same year and has spent most of the past 30 years developing physics models and detection/signal processing algorithms for a variety of submarine sensing technologies. He also has contributed to the development of tools for the early recognition of a biological attack, focusing on the evaluation of data sources and algorithms. Dr. Magruder's current assignment is as program manager for acoustics projects within the Ship Submersible Ballistic Nuclear (SSBN) Security Technology Program. **Amir H. Najmi** has been employed at APL since 1990. His main areas of research and expertise are spectral estimation and adaptive signal processing, wave propagation, and quantum theory of fields and particles. He has published in the most prestigious physics and geophysics journals on subjects as diverse as cosmological applications of quantum fields and seismic inversion for oil exploration. Since 2004, he has worked on algorithm developments for syndromic surveillance. He has published two papers on the applications of adaptive signal processing to public health surveillance. Dr. Najmi received a degree in mathematics (the Mathematical Tripos) from Cambridge University, and a doctorate (D.Phil.) in theoretical physics from Oxford University. **William Peter** received his B.S. in physics from the University of Southern California and his Ph.D. in physics from the University of California, Irvine. He has worked at Los Alamos National Laboratory in nuclear physics and high-performance computing and was one of the founding scientists of a medical imaging company. He has recently developed algorithms to speed up Monte Carlo calculations, improve clustering, and use information theory for data fusion. Dr. Peter came to APL in 2006 and now is working within the disease surveillance initiative on alert detection and data fusion. He also serves as the Editor-in-Chief of the *International Journal of Plasma Science and Engineering*. **Michael W. Thompson** earned a B.A. in music and a B.S. in applied physics from Brigham Young University in 1998. He subsequently earned an M.S. in physics from Brigham Young in 2000 doing computer simulations of nonlinear sound production in trombones. He continued his education at the Pennsylvania State University, where he earned a Ph.D. in acoustics in 2004 doing experimental studies of nonlinear acoustic streaming in thermoacoustic devices. In 2004, he joined the Acoustics and Electromagnetics Group (STX) in the National Security Technology Department at APL. He has since been working in the areas of disease surveillance and underwater acoustics. Dr. Thompson is a member of the Acoustical Society of America. For further information on the work reported here, contact Dr. Burkom. His e-mail address is howard.burkom@jhuapl.edu.

Howard S. Burkom

Yevgeniy Elbert

Steven F. Magruder

Amir H. Najmi

William Peter

Michael W. Thompson