

# Development, Adaptation, and Assessment of Alerting Algorithms for Biosurveillance

*Howard S. Burkom*

**T**he goal of APL's biosurveillance effort is to assist the public health community in the early recognition of disease outbreaks. ESSENCE II, the Electronic Surveillance System for the Early Notification of Community-based Epidemics, applies alerting algorithms to "anonymized" consumer data to give epidemiologists early cues to potential health threats in the National Capital Area. Raw data include traditional indicators such as hospital emergency room visits as well as nontraditional indicators such as physician office visits and less-specific, but potentially timelier, indicators such as sales of over-the-counter remedies. To improve the timeliness of alerting for disease outbreaks, we have adapted temporal and spatiotemporal algorithms from various disciplines, including signal processing, data mining, statistical process control, and epidemiology.

## INTRODUCTION

### The Biosurveillance Problem

The current global geopolitical climate and increased availability of affordable technology have combined to create concerns over the possibility of terrorist attacks using weapons of mass destruction. Various government agencies are funding defensive technology programs pertaining to nuclear, chemical, and biological threats. The biological threat is unique in that the covert release of a weaponized pathogen may precede casualties and other evidence of disease by days, allowing time for perpetrators to escape and leaving uncertainty as to when, where, and even whether an attack has occurred. Victims of a biological attack may be localized or widely scattered, and the population at risk may be difficult to identify, fostering the fear and panic that are the objectives of terrorism.

The weaponization of many diseases by well-funded offensive programs has been documented.<sup>1</sup> The symptomatology of these diseases has been studied, with a key finding<sup>2</sup> that their early presentation is usually characterized by influenza-like symptoms. An outbreak may include a prodromal period of mild symptoms before patients seek emergency care and before laboratory tests can identify a specific pathogen. If these mild symptoms were widespread, alerting algorithms using nontraditional data sources could give the public health system an early cue to respond to the outbreak.

This article focuses on algorithms developed to enable more rapid detection and characterization of biological attacks. Cost-benefit analyses, such as those

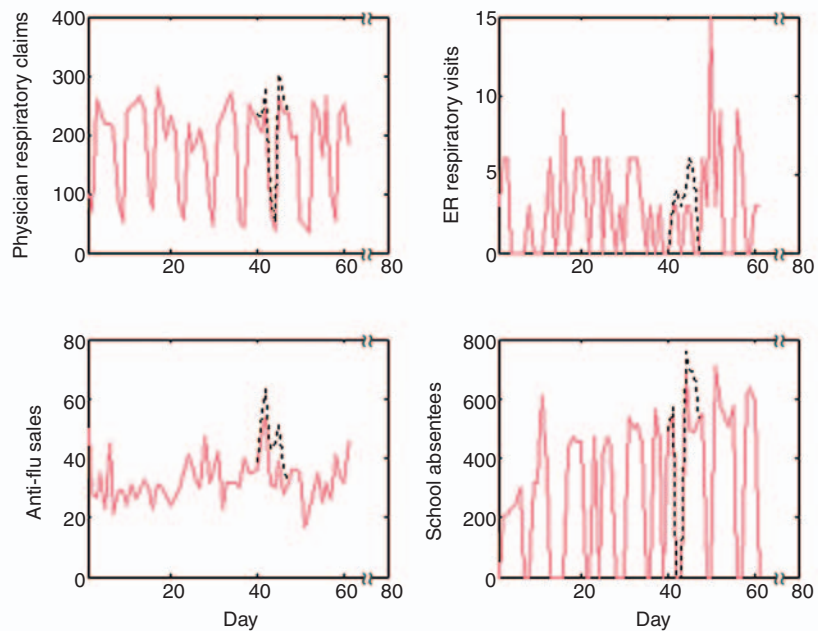
reported by Kaufmann et al.,<sup>3</sup> have shown potentially substantial savings in both human and economic terms if improved alerting methods can expedite a public health response to an outbreak by even a couple of days. The intended users of these algorithms and the systems that run them are epidemiologists of the public health infrastructure in government agencies. Therefore, an important requirement of the algorithms is that the expected number of alerts per time period be manageable by these public health users, given their resources for investigation.

Multiple algorithms are required to assess the probability of multiple types of threat, depending on the mode of dispersal of the pathogen, route of infection, and distribution of incubation periods of the resulting disease. Until very recently, only limited, highly specific data such as counts of emergency department visits for infectious diseases were used for surveillance. The raising of the stakes has stimulated intensive searches for improved alerting methods and for alternative data sources that might afford earlier warnings at tolerable false alarm rates.

### ESSENCE II Data Sources and Indicators

The article by Lombardo elsewhere in this issue describes the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE II) test bed that is integrating a growing set of civilian and military data sources within the National Capital Area. ESSENCE II includes traditional indicators such as hospital emergency room visits as well as nontraditional indicators such as physician office visits and less-specific but potentially timelier indicators including sales of over-the-counter (OTC) remedies and records of school and workplace absenteeism. All data used are anonymous, i.e., without identifying information, and include, for example, numbers and types of remedies purchased, number of patients with a particular diagnosis, or numbers of workers or students absent. Figure 1 illustrates the variation in time series behavior among the daily counts from these sources. Note the differences in scale, variability, and temporal behavior.

Additional issues include the potential for false alerts resulting from unrelated events, data-related behaviors following an outbreak, and known reporting



**Figure 1.** Variability of background and signal across data streams, including plausible effects of a disease outbreak (dashed curves represent the sum of cases without an outbreak plus cases attributable to an outbreak).

lags in data sources. Some events causing false alerts are unpredictable. In absenteeism data, school vacations and district-wide examinations can be modeled, but bomb scares and weather events cannot. Issues of behavior following an outbreak need to be modeled, such as how much and how soon an outbreak will affect a data source—the analog of propagation loss in detection theory. If there are 100 cases in a community outbreak, how many additional office visits and OTC purchases should we expect? Can knowledge of local demographics help us model these behaviors? Finally, if data history is available to model known reporting lags in counts of physician office visits and insurance claims, survival analysis methods may help to estimate actual data counts given reported levels.<sup>4</sup> The article by Magruder in this issue presents some of the data analysis efforts under way to examine these issues.

## ALERTING ALGORITHM METHODOLOGY

### Case Definition: Choosing the Appropriate Outcome Variable

Crucial to the utility of alerting algorithms is the selection of the outcome variable, i.e., the quantity to be measured among the growing volume of surveillance data. For medical data, ESSENCE II uses syndromic surveillance by monitoring counts of patient data from all military treatment facilities in the National Capital Area. Specifically, outpatient visits are monitored,

with diagnoses falling in any of seven syndrome groups chosen by physicians at the DoD Global Emerging Infections System (DoD-GEIS): Respiratory, Gastrointestinal, Fever, Dermatologic Infectious, Dermatologic Hemorrhagic, Neurologic, and Coma. A list of diagnostic codes from the International Classification of Diseases, Ninth Revision (ICD-9),<sup>5</sup> is defined for each of these syndrome groups, which have been gaining acceptance in the surveillance community.<sup>6</sup> ESSENCE increments the count for a syndrome group each time a diagnosis code falls in the corresponding list.

We have extended this practice to the monitoring of civilian claims and emergency room visits. We have also defined more specific outcome variables with lists of codes corresponding to the major symptoms associated with specific diseases such as anthrax. The case definition process is important for distinguishing the signal—the disease of interest—from the background, and we exclude some codes for that reason.

There is also a case definition process for nonmedical data. For example, counts of OTC sales are typically restricted to remedies for influenza or diarrhea, and studies are under way to further focus on products that would be popular purchases in the event of a given type of outbreak. Also, surveillance using school absenteeism data is typically restricted to younger students for whom absence is more indicative of illness.

### General Concepts for Alerting Algorithms

In recent years, algorithms for biosurveillance have been drawn from a variety of fields including epidemiology,<sup>7</sup> signal processing,<sup>8</sup> data mining,<sup>9</sup> and statistical process control.<sup>10</sup> Developers of these algorithms have sought to flag anomalies based on purely temporal behavior, space–time interaction, and unusual distributions of covariates such as an unexpected number of respiratory problems in a particular age stratum. These diverse methods share common underlying challenges. For example, are the data in the current test interval sufficiently different from expected counts to cause an alert? The data tested may be the latest elements in a single time series or a set of recent observations from disparate sources spread over the surveillance region. The expectation may be as simple as a scalar mean, but for adaptive detection performance, it is usually calculated from a recent baseline interval chosen to represent expected behavior. There are important factors in choosing this baseline:

- The choice of baseline length is a trade-off between modeling relationships among covariates and capturing recent trends.
- The end of the baseline may be chosen to leave a gap, or “guardband,” before the test interval to exclude the early part of a true outbreak from the data used for expectation.

- Data in the baseline may be smoothed, weighted, or even selectively censored if outliers irrelevant to the alerting process can be identified.

Once the test and baseline data are chosen, we compute a test statistic and compare it to an alerting threshold, which may also be adaptively computed. Because of the complexity of the test statistic for a number of algorithms of recent interest,<sup>7,9</sup> the data in the baseline are repeatedly randomized in a Monte Carlo process to determine the alerting threshold empirically.

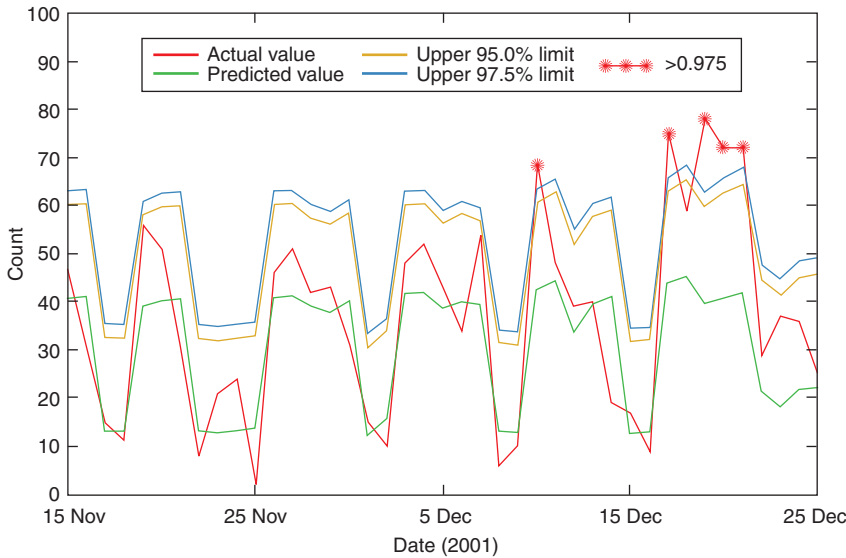
### Purely Temporal Methods

Purely temporal alerting algorithms seek anomalies in single or multiple time series without location or distance information. We use two basic temporal approaches: regression-based behavior modeling and adaptive process control charts.

The modeling approach currently implemented in ESSENCE II includes trend; categorical variables for weekends, holidays, and post-holidays; and adaptive autoregressive terms to account for serial correlation. A sliding 28-day baseline is used to compute regression coefficients and an updated standard error for the residuals, i.e., the differences between observed and predicted values. Coefficients are used to make predictions, and the test statistic is the current-day standardized residual. Figure 2 illustrates the modeling of counts of influenza-like illness claims from military data sources in a Maryland county. Note the weekly pattern in the red curve indicating counts of claims. The green curve shows the modeled counts, and asterisks indicate alerts at the  $3\sigma$  level based on the assumption that residuals are normally distributed. Regression methods using additional covariates are under development.

Sometimes the data counts of interest are not readily modeled, as when data history is short or counts are sparse. Emergency room admissions data provide a good example, especially when the counts are taken from a small geographic region. In such cases we use adapted process control methods, which generally operate on some measure of how counts in the test interval vary from the baseline mean. For example, the Early Aberration Reporting System (EARS) algorithms developed by the Centers for Disease Control and Prevention (CDC) are used by many local health departments across the United States. Although these algorithms use only a 7-day baseline, they have performed well in comparisons with far more complex methods that use models based on long data history.<sup>11</sup> ESSENCE II includes the EARS algorithms among other process control techniques.

For a simple illustration, we present a method adapted from the exponential weighted moving average (EWMA) chart given by Ryan.<sup>12</sup> Let  $X_t$  be a time series of values,  $t = 1, \dots, n$ , and for some smoothing constant  $\omega$ ,  $0 < \omega < 1$ , form the smoothed value  $Y_t$  by



**Figure 2.** Autoregressive modeling of syndromic data for influenza-like illness from a large county in Maryland. Data are shown in red, predictions in green; indicated confidence limits reflect probabilities assuming Gaussian-distributed residual values.

$$Y_1 = X_1,$$

and

$$Y_t = \omega * X_t + (1 - \omega) * Y_{t-1} .$$

Note that  $Y_t$  is a function of the current count  $X_t$  here; in some implementations  $X_t$  is replaced by  $X_{t-1}$  to obtain predicted values. The  $Y_t$  give a weighted moving average of the observations with increased emphasis on more recent observations; a larger  $\omega$  increases this emphasis, while a small  $\omega$  spreads the weighting more among past observations.

Now let  $\mu_t$  be the mean and  $\sigma_t$  the standard deviation of the data in the current baseline. If the test statistic  $(Y_t - \mu_t) / \{\sigma_t[\omega / (2 - \omega)]^{1/2}\}$  exceeds the threshold value, often set at 3 in process control applications, an alert is flagged. A 2-day guardband is used for the baseline computations to avoid missing a gradual buildup over several days, and the smoothed value  $Y_t$  is reset to the next data value after an alert to avoid residual alerting after very large values.

Figure 3 illustrates a spreadsheet method used for the parametric testing of algorithms. The characteristic parameters and the baseline length, guardband, and threshold may be adjusted and algorithm performance examined using arbitrary

data streams with outbreak “signals” injected. (More statistical methods for assessing algorithm performance are discussed in “Assessment Using Monte Carlo Trials” later in this article.)

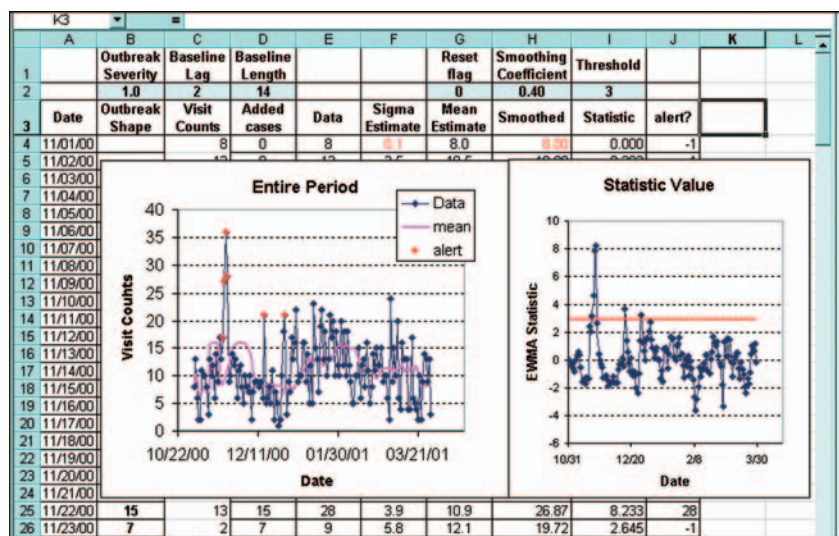
### Time-Domain Matched-Filter Approach

For data from multiple sources, we have implemented an adaptive, time-domain matched filter. The adaptive matched filter was developed in the radar community as an optimal detector in the presence of Gaussian noise and has been used widely in a variety of noise environments as discussed in Cook and Bernfeld.<sup>13</sup> This technique is appropriate for problems in which time variation of the signal is

known sufficiently to model the signal as a mathematical replica.

The matched filter is designed to find signals that match the expected replica signal and reject signals or noise that is unlike the replica. The usual procedure effectively takes the normalized inner product of successive segments of an input data stream with the replica. Thresholds are then applied to these successive products to make detection decisions.

There were two reasons for adopting the adaptive matched-filter approach. First, the ramping and peaking of public health data at the onset of an outbreak indicate a time-varying signal, and this signal may be



**Figure 3.** Parametric algorithm analysis: (left) smoothing algorithm applied to visit counts for the entire period, and (right) test statistic showing when the value has exceeded the threshold of 3.

modeled as an epidemic curve as noted later in the “Modeling the Signal” section. Second, an adaptive matched filter can handle disparate characteristics of noise from different data sources. An optimal detector must consider the noise background as well as the signal model. Data channels that have significant noise fluctuations imitating the desired signal and causing false alarms should be suppressed, while channels with low noise should be emphasized for increased sensitivity. The adaptive matched filter estimates the noise in each channel with covariance matrices computed using neighboring channels.

For the implementation of the matched filter, suppose that the filter extends over  $N$  days of data and that  $\mathbf{X}_i$  is the vector of residual data at day  $i$ . Typically, the first  $J$  elements of  $\mathbf{X}_i$  are residuals derived from school absentee rates  $1, \dots, J$  for that day, the next  $K$  elements are from OTC sales at stores  $1, \dots, K$ , etc. Let  $\mathbf{C}_i$  be the estimated covariance matrix of  $\mathbf{X}_i$ , and let  $\mathbf{r}$  be a replica vector of modeled effects of the outbreak on the data. The normalized replica is then  $\mathbf{M}_i = \mathbf{r}/(\mathbf{r}\mathbf{C}_i\mathbf{r}')^{1/2}$ . The adaptive matched filter is given by

$$y = \sum_{i=1}^N \mathbf{M}_i^T \mathbf{C}_i^{-1} \mathbf{X}_i.$$

The adaptive matched-filter statistic, also found in McDonough and Whalen,<sup>14</sup> is  $\mathbf{r}\mathbf{M}_i^{-1}\mathbf{X}_i$ . More detail on this approach is given in Burkom et al.<sup>15</sup>

### Spatiotemporal Methods Based on Scan Statistics

In the general context of surveillance of a large region, we monitor the separate subregional time series corresponding to each data source. The series of counts for OTC sales may be binned by store location; for claims, by facility or patient zip code; for absenteeism, by school or work site. Use of the spatial dimension offers two advantages: greater sensitivity to small increases in counts and potential inferences from spatial relationships among subregions. The granularity of these subregions and the outcome variable chosen dictate the applicable algorithms; finer subdivisions and smaller counts mean greater sensitivity but less structure in the data for modeling approaches.

Several spatial approaches have been tried, including the application of standard contingency tables. For this approach we have replaced dichotomies of exposed-versus-unexposed and cases-versus-controls by current-versus-background and inside-versus-outside the region of interest. Statistical significance indicates an association of the region of interest with the measurement time window rather than an association of disease with exposure. The estimation of “normal” background counts is an important step in this process and indeed in any detection scheme. Depending on the specific data and spatial scale, we have used regional population, number

eligible for services per region, and recent baseline values.

A spatial approach that has proven effective in the biosurveillance context is the scan statistic. The version presented by Kulldorff,<sup>16</sup> referred to later as the Kulldorff statistic, has been widely used, particularly in cancer epidemiology. Our early efforts employed the SaTScan implementation, which is downloadable from the National Cancer Institute’s Web site.<sup>17</sup>

### Adapting the Kulldorff Scan Statistic

We briefly cast the SaTScan approach in the context of the general surveillance problem. (A rigorous presentation and fuller discussion are given elsewhere.<sup>7,16</sup>)

1. Subdivide the surveillance region into subregions  $j = 1, \dots, J$  of which centroids or other representative points are used for cluster analysis.
2. For a given data source, tabulate observed counts  $O_j$  for each subregion—typically the number of outpatient visits with a diagnosis in a specified syndrome, the count of sales of anti-flu remedies, etc.
3. Given the sum  $N$  of all subregional counts, calculate the expected counts  $E_i$  for each subregion. In the conventional use of SaTScan, these counts are assumed to be proportional to subregion populations. Accuracy and stability of the expected spatial count distribution are essential to avoid computing spurious clusters that can mask the case groups of interest. Since our data streams are typically not population-based, we use modeling or data history with a 2- to 4-week baseline to estimate this distribution.
4. The hypothesis is that for some subset  $J1$  of the  $J$  subregions, the probability that an outbreak has occurred is  $p$ , while the probability for subregions outside of  $J1$  is some  $q < p$ . The null hypothesis, then, is that  $p = q$  for all subsets  $J1$  of  $J$ .
5. Candidate clusters are formed by taking families of circles centered at each of a set of grid points—often taken as the full set of subregion centroids. A candidate cluster is defined as those subregions  $j$  whose centroids lie in the associated circle. For each grid point, candidate cluster sizes range from a single subregion up to a preset maximum fraction of the total count.
6. For each candidate cluster  $J1$ , under the assumption that cases are Poisson-distributed in space, the likelihood ratio for the clustering hypothesis is then

$$LR(J1) \equiv (O1/E1)^{O1} * [(N - O1)/(N - E1)]^{(N - O1)},$$

where  $O1$  and  $E1$  are the observed and expected counts summed from subregions in  $J1$ , respectively, and  $N$  is the sum of counts in all subregions.

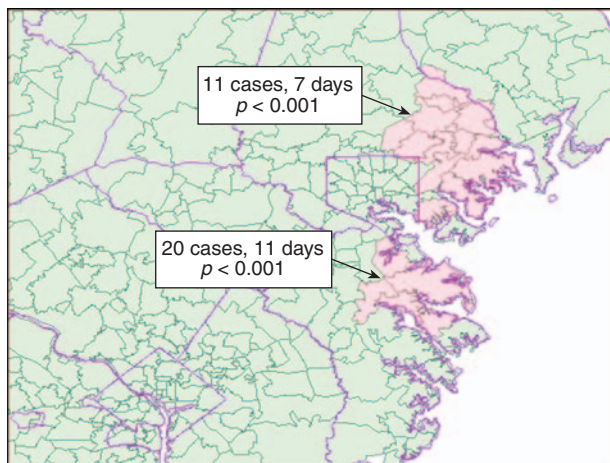
7. The maximal cluster is then taken to be the set  $J1^*$  of subregions corresponding to the circle with the

- maximum likelihood ratio over all grid centers and all circles.
8. A  $p$ -value estimate for the statistical significance of this cluster is determined empirically by ranking the value of  $LR(J1^*)$  among other maximum likelihood ratios, each calculated similarly from a random sample of the  $N$  cases based on the expected spatial distribution.
  9. Once a set of subregions is associated with a maximal cluster, secondary clusters are chosen and assigned significance levels from the successively remaining subregions.

The Kulldorff statistic has several advantages over other spatial methods.<sup>18</sup> It yields both cluster locations and significance levels and avoids preselection bias and multiple testing effects; can catch outbreaks scattered among neighboring regions that more sophisticated modeling methods may miss; automatically adjusts for temporal nationwide patterns, i.e., a seasonal increase proportionately affecting all subregions will not affect the statistic; and can complement temporal modeling methods by providing candidate clusters for more detailed computations.

An early application to ESSENCE data was a retrospective test to detect a known scarlet fever outbreak east of Baltimore, Maryland, in the winter of 2001 using electronic military and civilian patient data. The only data characteristics used for this study were the date of patient visit, assigned ICD-9 codes, and patient home zip code. For the outcome variable, we used daily counts in each zip code of claims whose diagnoses included codes 034 (Scarlet Fever) or 034.1 (Strep Throat Due to Scarlet Fever).

Figure 4 indicates clusters of cases found on 18 January 2001 along with their empirically determined significance levels. Such plots allow the medical analyst



**Figure 4.** Clusters (shown in pink) found using spatial scan statistics (SaTScan software) on 18 January 2001 in the National Capital Area for a scarlet fever outbreak study.

to follow the spread of an outbreak. More recently, we have used groups of diagnosis codes included in the ESSENCE II syndrome groups and selected subgroups to seek and track clusters of endemic disease and to look for unexpected outbreaks. We have found transient clusters of influenza-like illness during cold season and occasional small outbreaks of other nonreportable illnesses.

Recent efforts have focused on finding clusters of space–time interaction using multiple data sources without counts from high-variance sources that mask moderate signals in sources of smaller scale and variance. The fusion of such disparate sources is difficult because counts cannot be scaled by their variance in the conventional SaTScan approach. A stratified scan statistic<sup>19</sup> has shown promise for finding signals in both low- and high-variance sources without losing power to detect faint signals distributed throughout all sources.

## ALGORITHM PERFORMANCE ASSESSMENT

### Modeling the Signal

Any assessment of the performance of a detection method requires knowledge of the signal of interest. We consider the signal to be the number of additional data counts (syndromic cases, OTC sales, absentees, etc.) attributable to a point-source outbreak—the feared result of a terrorist release of a pathogen in a public place—on each day after exposure. Lacking this specific knowledge, we assume that the “data epicurve” for each source is proportional to the number of new symptomatic cases each day, excluding weekends, with obvious adjustments such as cumulative attributable school absences. The signal problem is then to estimate the number of new symptomatic cases by day after exposure.

It is difficult to use authentic outbreaks to make such estimates for two reasons: (1) outbreaks are not readily identifiable in datasets of interest, and (2) our objective is to assess early reporting capability, so we need to know when an outbreak begins, ideally to within a day. We therefore get our epicurve estimates by simulation.

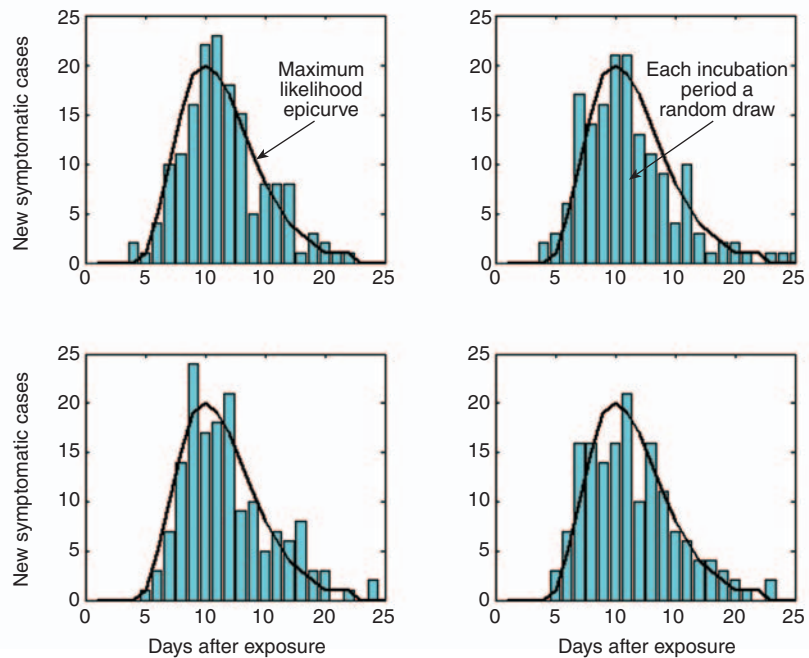
For an epicurve model, we use the lognormal distribution first discussed by Sartwell<sup>20</sup> in 1949 and widely used since then, as noted in the review sections of Philippe<sup>21</sup> and Armenian and Lilienfeld.<sup>22</sup> From an analysis of the incubation periods of 17 infectious diseases, Sartwell observed that incubation period distribution tends to be lognormal, with distribution parameters dependent on the disease agent and route of infection. To model a specific disease, we infer the values of lognormal parameters from histograms of incubation period data as in Meselson et al.<sup>23</sup> or from less-specific medical literature such as the U.S. Army Medical Research Institute’s “blue book.”<sup>24</sup> Given these parameters and a hypothetical number of cases, we may produce a model epicurve

of primary symptomatic cases—all cases if the disease is not communicable. The solid curves in Fig. 5 show such an epicurve for parameters derived from the smallpox outbreak data used in Hill.<sup>25</sup>

### Assessment Using Monte Carlo Trials

To test the performance of an algorithm on a given set of data, we construct many realizations of a chosen epicurve model for a fixed number  $N$  of symptomatic cases. To make the trials challenging for the algorithms, we choose the peak value of the epicurve to be a multiple (usually 2 or 3) of the standard deviation of the data, and then  $N$  is this peak value divided by the maximum of the lognormal density function. Having set the epicurve shape and the number of cases  $N$ , we generate the set of incubation periods with a set of  $N$  random lognormal draws and round each to the nearest day, just as each actual incubation period could be seen as a random draw from distributions of dosage and susceptibility. The number of cases to add for each day is then the number of draws rounded to that day. Figure 5 shows four sample stochastic epicurves generated with this method.

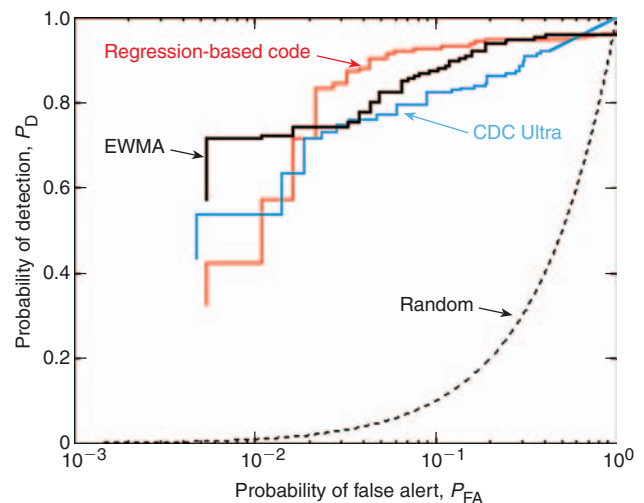
We conduct each trial by adding a stochastic epicurve to a randomly chosen start day in the test data. The start day is usually chosen beyond some warm-up interval required by the algorithm. We then apply the algorithm to the modified data and note whether it flags an anomaly during the outbreak day(s) of interest for a given threshold. The empirical detection probability  $P_D$  for this threshold is the fraction of all trials for which the outbreak is detected in this sense. The false alert probability  $P_{FA}$  is the fraction of days with no added counts on which the algorithm flags an anomaly for the threshold. We obtain a receiver operating characteristic (ROC) curve for the chosen data, outbreak shape, and outbreak severity by plotting  $P_D$  versus  $P_{FA}$  for a set of values of the threshold. Figure 6 shows a set of these curves comparing the CDC Ultra algorithm, an EWMA method, and a regression-based code using counts of diagnoses in the Gastrointestinal syndrome group. The preferred algorithm in these comparisons often depends on the allowable false alert level. We use this ROC assessment method to compare algorithms, to choose optimal algorithm parameters, and to estimate minimal detectable signals.



**Figure 5.** Four randomly chosen incubation period distributions used for signal representation from the Sartwell model<sup>20</sup> (solid curves) with parameters  $\zeta=2.4$ ,  $\sigma=0.3$ , and cases at peak = 20. The x axis gives incubation time in days.

### SUMMARY

This article has presented an overview of the methodology of algorithms used for biosurveillance. Such surveillance presents new technical challenges. Current research is dealing with a variety of data sources offering



**Figure 6.** Algorithm performance analysis with receiver operating characteristic (ROC) curves based on simulated outbreaks. Background data were Gastrointestinal syndrome counts, simulated outbreak peaks were 3 times the data standard deviation, and detection probabilities were tabulated 2 days before the signal peak. Best performance is for the highest probability of detection along with lowest (or allowable) probability of false alert.

many different noise backgrounds. The signal of interest is the effect of a hypothetical outbreak of disease on these data sources; however, among the many possible threat scenarios, this effect is poorly understood. Modeling methods and statistical tests from several disciplines are being adapted and tested as public health anomaly detectors. We are working with single and multiple data streams, purely temporal and spatiotemporal methods, and a variety of data fusion approaches to determine efficient combinations of data sources and indicators and algorithms for monitoring them. This is part of the APL effort to help mitigate the damage caused by a possible bioterrorist attack. Continuing research on these algorithms is proceeding side by side with the data analysis effort in an overall approach that is highly data-driven.

## REFERENCES

- <sup>1</sup>Alibek, K., *Biohazard*, Random House, New York (1999).
- <sup>2</sup>Franz, D. R., Jahrling, P. B., Friedlander, A. M., McClain, D. J., Hoover, D. L., et al., "Clinical Recognition and Management of Patients Exposed to Biological Warfare Agents," *JAMA* **278**(5), 399–411 (6 Aug 1997).
- <sup>3</sup>Kaufmann, A. F., Meltzer, M. I., and Schmid, G. P., "The Economic Impact of a Bioterrorist Attack: Are Prevention and Postattack Intervention Programs Justifiable?" *EID* **3**(2), 83–94 (Apr–Jun 1997).
- <sup>4</sup>Brookmeyer, R., and Gail, M. H., "Statistical Issues in Surveillance of AIDS Incidence," Chap. 7 in *AIDS Epidemiology: A Quantitative Approach*, Oxford University Press, New York (1994).
- <sup>5</sup>*Diagnostic Coding Essentials*, Ingenix Publishing Group, Salt Lake City, UT (2001).
- <sup>6</sup>Lazarus, R., Kleinman, K., Dashevsky, I., DeMaria, A., and Platt, R., "Use of Automated Ambulatory-Care Encounter Records for Detection of Acute Illness Clusters, Including Potential Bioterrorism Events," *EID* **8**(8), 753–760 (Aug 2002).
- <sup>7</sup>Kulldorff, M., "Spatial Scan Statistics: Models, Calculations, and Applications," in *Scan Statistics and Applications*, J. Glaz and N. Balakrishnan (eds.), Birkhauser, Boston, MA, pp. 303–322 (1999).
- <sup>8</sup>Goldenberg, A., Shmueli, G., Caruana, R., and Fienberg, S., "Early Statistical Detection of Anthrax Outbreaks by Tracking Over-the-Counter Medication Sales," *Proc. Nat. Acad. Sci. USA* **99**(8), 5237–5240 (16 Apr 2002).
- <sup>9</sup>Wong, W.-K., Moore, A. W., Cooper, G., and Wagner, M., "Rule-based Anomaly Pattern Detection for Detecting Disease Outbreaks," in *Proc. 18th Nat. Conf. on Artificial Intelligence*, MIT Press, pp. 217–223 (2002).
- <sup>10</sup>Ngo, L., Tager, I. B., and Hadley, D., "Application of Exponential Smoothing for Nosocomial Infection Surveillance," *Am. J. Epidemiol.* **143**(6), 637–647 (1996).
- <sup>11</sup>Hutwagner, L., Thompson, W., Groseclose, S., and Williamson, G. D., "An Evaluation of Alternative Methods for Detecting Aberrations in Public Health Surveillance Data," in *2000 Proc. Biometrics Section, Am. Statistical Assoc.*, pp. 82–85 (Aug 2000).
- <sup>12</sup>Ryan, T. P., *Statistical Methods for Quality Improvement*, John Wiley & Sons, Inc., New York (1989).
- <sup>13</sup>Cook, C. E., and Bernfeld, M., *Radar Signals, An Introduction to Theory and Application*, Academic Press Inc., Orlando, FL (1967).
- <sup>14</sup>McDonough, R. N., and Whalen, A. D., "The General Discrete Matched Filter," Section 6.6 in *Detection of Signals in Noise*, Academic Press, San Diego, CA (1995).
- <sup>15</sup>Burkom, H. S., Lombardo, J. S., Newhall, B. K., Pineda, F. J., and Chotani, R. A., *Automated Alerting for Bioterrorism Using Autonomous Agents*, STN-00-373, JHU/APL, Laurel, MD (Mar 2001).
- <sup>16</sup>Kulldorff, M., "A Spatial Scan Statistic," in *Communications in Statistics—Theory and Methods* **26**, pp. 1481–1496 (1999).
- <sup>17</sup>SaTScan™ Version 2, developed by M. Kulldorff, available at <http://srab.concer.gov/satscan/> (accessed 28 Aug 2003).
- <sup>18</sup>Kulldorff, M., and Nagarwalla, N., "Spatial Disease Clusters: Detection and Inference," *Stat. Med.* **14**, 799–810 (1995).
- <sup>19</sup>Burkom, H. S., "Biosurveillance Applying Scan Statistics with Multiple, Disparate Data Sources," *J. Urban Health, Proc. 2002 Nat. Syndromic Surveillance Conf.* **80**(2), Suppl. 1, pp. i57–i65. (Apr 2003).
- <sup>20</sup>Sartwell, P. E., "The Distribution of Incubation Periods of Infectious Disease," *Am. J. Hyg.* **51**, 310–318 (1950).
- <sup>21</sup>Philippe, P., "Sartwell's Incubation Period Model Revisited in the Light of Dynamic Modeling," *J. Clin. Epidemiol.* **47**(4), 419–433 (1994).
- <sup>22</sup>Armenian, H. K., and Lilienfeld, A. M., "Incubation Period of Disease," *Epidemiol. Rev.* **5**, 1–15 (1983).
- <sup>23</sup>Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., et al., "The Sverdlovsk Anthrax Outbreak of 1979," *Science* **266**, 1202–1208 (1994).
- <sup>24</sup>*Medical Management of Biological Casualties*, U.S. Army Medical Research Inst. of Infectious Diseases, Ft. Detrick, MD (Sep 2000).
- <sup>25</sup>Hill, B. R., "The Three-Point Lognormal Distribution and Bayesian Analysis of a Point-Source Epidemic," *J. Am. Stat. Assoc.* **58**(301), 72–84 (1963).

ACKNOWLEDGMENTS: This research is conducted in full compliance with the Health Insurance Portability and Accountability Act. It is sponsored by the Defense Advanced Research Projects Agency and managed by the Naval Sea Systems Command under contract N00024-98-D-8124. I acknowledge Bruce Newhall, Anshu Saksena, and Fernando Pineda for their initial scenario and matched-filter work; Yevgeniy Elbert of the Walter Reed Army Institute for Research for the application of autoregressive modeling; and Farzad Mostashari of the New York City Health Department for consulting and for introducing scan statistics to this program.

## THE AUTHOR



HOWARD S. BURKOM received a B.S. degree in mathematics from Lehigh University in 1970, an M.S. degree in mathematics from the University of Illinois/Champaign in 1971, and a Ph.D. in geometric topology from the University of Illinois in 1978. Dr. Burkom has worked at APL since 1979 in underwater acoustics, tactical oceanography, and public health surveillance. Recent projects have involved the application of alerting algorithms from the fields of epidemiology, biostatistics, data mining, and statistical process control to the monitoring of clinical and nontraditional data sources for outbreaks of disease. He now serves as manager of the Anomaly Discovery effort in the Laboratory's National Security Technology Department. His e-mail address is [howard.burkom@jhuapl.edu](mailto:howard.burkom@jhuapl.edu).