Analytic Biosurveillance Methods for Resource-Limited Settings

Howard S. Burkom, Yevgeniy A. Elbert, Jacqueline S. Coberly, John Mark Velasco, Enrique A. Tayag, and Vito G. Roque Jr.

Public health surveillance faces many challenges in geographic regions lacking modern technology and infrastructure. This article addresses the role of analytic methods in such regions and evaluates temporal alerting algorithms using both authentic and simulated data sets. Evaluation analyses give the technical background for the statistical methods provided by the Johns Hopkins University Applied Physics Laboratory (APL) Suite for Automated Global Electronic bioSurveillance (SAGES), a collection of modular, open-source software tools to enable electronic surveillance in resource-limited settings. Included in the evaluation are only those statistical methods that are broadly applicable to multiple evolving-background time-series behaviors with limited data history. Multiple detection performance measures are defined, and a practical means of combining them is applied to recommend preferred alerting methods for common scenarios. Effective usage of these methods is discussed in the context of routine health-monitoring operations.

INTRODUCTION

Background

The 21st century has seen advances in many aspects of global disease surveillance.[1] These advances have been driven by heightened concerns over perceived threats to public health both from natural pathogens and from bioterrorism. These concerns have led to mandated improvements at the international level, through revision of International Health Regulations of the World Health Organization,[2] and also in the United States at the national level.[3]

Particular concern exists for surveillance in resource-limited settings (i.e., areas with limited access to medical care; inadequate or no laboratory diagnostic capability; insufficient numbers of first responders, care providers, and public health workers; and sometimes deficiencies in fundamental hygienic needs such as clean drinking water).[4,5] As a result of these issues, such regions are vulnerable to outbreaks of diseases, such as cholera and typhoid fever, that are not seen in advantaged settings.

SAGES Program at APL: Mission, History, Status

The Johns Hopkins University Applied Physics Laboratory (APL) has contributed advances in electronic disease surveillance since the late 1990s,[6] before the surge of development stimulated by the terrorist attacks of 2001. SAGES (Suite for Automated Global Electronic bioSurveillance) is a collaboration between APL and the US Armed Forces Health Surveillance Center to extend these advances to resource-limited settings. SAGES is an open-source software tool set for data collection, analysis, visualization, and reporting.[7] These tools were designed to maintain and extend established user-driven features of ESSENCE (Electronic Surveillance System for the Early Notification of Community-based Epidemics).[6] The tools were designed to meet a range of institutional needs and capabilities and for convenient integration with local health-monitoring tools. The purpose of this article is to improve the SAGES evaluation/visualization tools by identifying and tuning statistical alerting methods for given contexts. For example, given the amount of historical data available, whether monitoring is done daily or weekly, and whether data are sparse or rich with cyclic patterns, which alerting methods should be used?

Alerting Methods: Principles and Current Objective

The value of statistical alerting methods in a syndromic surveillance system is for detection of statistical anomalies, not detection of actual health events whose confirmation requires definitive evidence that is not immediately available to most SAGES users. In such a system, an alert is signaled when the output of an algorithm for monitoring a data stream crosses a threshold indicating behavior that is statistically aberrant, or too far from expected values to be plausible from random variation alone. The anomaly alerts, especially in combination with other evidence, are useful to prompt investigation of true health events, but the alerts have other causes, including batched data reports, changes in data participation, and changes in diagnosis coding. This article provides the basis for the alerting methods and chosen parameters supplied through the SAGES website[7] as of early 2014, with guidance for effective usage. This guidance does not require sophisticated or time-consuming analysis from SAGES users, who range from part-time technicians to medically trained epidemiologists with varying backgrounds and levels of availability.

METHODS

Selection of Candidate Alerting Methods

The data available to current and near-term SAGES users restricted this project and the initial SAGES open-source alerting methods to algorithms for a single time series derived by aggregation of select clinical data on individuals. For example, a typical input time series is the succession of daily or weekly counts of medical encounter records whose chief complaint field contains words related to febrile illness, such as "temperature," "fever," and "feverish." Candidate alerting methods were algorithms that could flag relevant target signals in the data at manageable background alarm rates. The term *background alarm* is used here in place of *false alarm* because false positives are difficult to verify in authentic surveillance data; indeed, given the practical constraints of public health response, many alerts are not investigated at all. The section called *Target Signals for Aberration Detection* explains how true positives were determined in the two phases of this work.

It is assumed that SAGES users have access only to selected clinical report counts, not to exogenous clinical variables, nor to nonclinical information such as environmental data. Candidate methods may not assume more than a few months of data history, because even in situations when quality data are available for multiple successive years, the older data may not be useful for training or baseline determination because of changes in data provider participation, information systems, or diagnosis coding. Another requirement is that the methods be easily implemented and maintained without assuming future availability of statistical expertise for tuning or model refitting. Implementation and routine usage on health monitoring systems rules out methods that require excessive time to calculate baselines and produce results for input data streams that may be improvised. Transparency is also an essential requirement for SAGES alerting algorithms; the user need not understand the underlying mathematical detail, but the basic concept should be clear enough that the SAGES user can see why an alert is indicated. Many methods noted in recent survey articles[8] do not meet all of these requirements.

For these reasons, the initial set of SAGES alerting methods was restricted to adaptive versions of the control charts long used in the statistical quality assurance community.[9] Adaptive features, noted in the descriptions below, were considered essential for alerting that is robust relative to common data quality issues such as data dropouts and abrupt changes in the background mean for prompt recovery of sensitivity after a large authentic or artifactual spike. Future enhancements will modify the provided alerting methods according to evolving needs and capabilities of SAGES users.

Candidate Alerting Methods with Descriptions

This section briefly describes four chosen alerting methods, denoted Z-score_SAGES, EWMA_SAGES, CuSUM_SAGES, and GS_SAGES. Each is intended to alert when the excess of recent time series values above the baseline expectation is statistically significant, indicating the possibility of an outbreak. These tests ignore

anomalously low values. In each method, the modifications below, detailed in a previous issue of this journal,[10] are implemented where applicable to account for common characteristics and challenges of the surveillance data environment:

- **Evolving data streams:** Candidate methods use a sliding, fixed-length baseline to calculate the mean and standard error, in place of values derived from phase I analysis in industrial control charts. In the latter context, in-control data behavior is typically stable, and data-generating processes can be stopped and adjusted when the data go out of control.
- **Accommodation of sparse or vanishing baselines:** Each method uses a minimum baseline standard deviation to avoid excessive statistic values and to enable the use of these methods for sparse data streams.
- **Non-Gaussian distribution of time-series data:** A correction to computed *p*-values is applied to account for the fact that count distributions are typically Poisson rather than Gaussian.
- **Robustness to data dropouts:** The methods test for historically implausible strings of zeros and reset to avoid prolonged, excessive alerting when data reporting resumes.

Based on the issues described above and on the practical requirement that methods must produce sensible alerting for time series formed from ad hoc queries without noticeable response delays, the following methods were chosen for comparison.

Z-score_SAGES

This method implements a standard control chart, an X-bar chart that is a generalization of the EARS C2 algorithm,[11] globally the most widely used alerting method for biosurveillance. Like the C2 method, it uses a sliding baseline with a fixed buffer between the test period and baseline. The test statistic at time step t is then the Z-score z_t,

$$z_t = (x_t - \overline{X}_t)/\hat{s}_t, \tag{1}$$

where x_t is the current time series element, $\overline{X}_t$ is the mean of the time series over the current baseline, and $\hat{s}_t$ is the baseline standard deviation. Numerous global implementations of C2 use a 7-day baseline, 2-day guard band, and a fixed alerting threshold of 3 standard deviations above the mean. ESSENCE and other systems have expanded the baseline to 28 days to achieve more stable alerting behavior. In the current study, the baseline, guard band, and threshold are varied for optimal detection performance. A *p*-value threshold is derived as a lookup of the Z-score value using the Student's t distribution with degrees of freedom equal to the baseline length – 1.

EWMA_SAGES

The exponential weighted moving average (EWMA) method evaluated for SAGES replaces the current observed value x_t in Eq. 1 with the recursively weighted average,

$$E_t = \omega\, x_t + (1-\omega)E_{t-1}, \tag{2}$$

for a fixed smoothing constant ω, $0<\omega<1$, that expresses how the weight of tested observations is distributed backward in time. For a value of ω near 1, only the most recent values influence this average, whereas reducing the value of ω increases the influence of older values.[12] Statistical corrections for the weighted averaging are applied to the standard error and threshold calculations.[10] This modification adds sensitivity to the gradual signals that form the data signature of many outbreaks of interest, but it reduces sensitivity to single spikes. For the current evaluation, combinations of the baseline value, guard band, smoothing constant, and alerting threshold were tested to seek the best detection performance.

CuSUM_SAGES

Like the EWMA method, the CuSUM chart has also been shown to be timelier than the X-bar chart at detecting small mean shifts and gradual signals.[13] Applications for biosurveillance have focused on aberrations above the baseline expectation.[14] The CuSUM_SAGES implementation uses a sliding baseline as in the Z-score_SAGES method and recursively calculates an upper sum $S_{H,t}$ of scaled differences z_t of x_t above the baseline mean estimate $\overline{X}_t$:

$$S_{H,t} = \max(0, z_t - \mathrm{k} + S_{H,t-1}). \tag{3}$$

In this expression, only differences of the observation x_t above ($\overline{X}_t$ plus k standard deviations) are added to the running test sum, while smaller differences are ignored. Equation 3 assures that the upper sum is nonnegative, and the method alerts if $S_{H,t}$ exceeds a computed threshold. The initial value $S_{H,0}$ of this statistic is set at half the alerting threshold to enable prompt alerting as in the Fast-Impulse-Response CuSUM,[15] and it is reset to this value after an alert to avoid persistent, unwanted alerts because of an extremely high (possibly erroneous) value, while still maintaining sensitivity.

Although many authors have found the detection performance of the CuSUM similar to that of the EWMA, the methods are substantially different. A CuSUM chart does not use the strict time-based weighting of past observations but is influenced only by those observations with scaled baseline exceedance above a fixed level. For

this reason, the CuSUM threshold is usually determined empirically,[9] and the CuSUM_SAGES version derives *p*-values from lookup tables calculated by running the algorithm on simulated time series of length 100,000.

GS_SAGES

This method was adapted[16] for SAGES user groups wishing to monitor daily count data, such as counts of selected clinic reports. In many SAGES settings, only count data are available, with no catchment or population-at-risk data to allow estimation of incidence rates. The main utility of this method is for monitoring daily time series with systematic day-of-week effects, including any regular weekly pattern, which the method can infer and progressively modify. For monitoring with controlled bias, statistical alerting algorithms must adjust to such known systematic data behaviors. The GS_SAGES algorithm is robust to common situations such as clinic closings on weekends or only on a particular day of the week and also on known calendar holidays. The algorithm requires only a couple of months of representative startup data. It is useful for series of counts that are not too sparse in the sense that an overall median count of at least three reports per day is preferred. For median counts closer to zero (i.e., no reports on the majority of days), a simpler adaptive method based on the EWMA or CuSUM control chart is recommended (see the section *Results Using Simulated Syndromic Data*).

The GS_SAGES algorithm applies generalized exponential smoothing for rapid adjustment to short-term trends and weekly patterns. It has been chosen in place of regression modeling as used in other systems[10] because conventional least-squares regression has been shown vulnerable to large errors when short-term trends affect the data.[17] GS_SAGES uses a prediction based on recursive smoothing equations given below, not on regression or any other global model. These equations can adapt to local changes in the mean value of the observed counts, as in a conventional EWMA chart, but the smoothing is generalized so that the prediction can also adapt to changes in the trend and in the weekly pattern.

For the smoothing equations, let α, β, γ denote smoothing coefficients for updating terms corresponding to the level, trend, and seasonality, respectively, and let s be the length of the season in the data. If y_t is the observed count on day t, then terms for level m_t, trend b_t, and seasonality c_t are updated as with the following equations:

$$\text{Level: } m_t = \alpha \frac{y_t}{c_{t-s}} + (1-\alpha)(m_{t-1} + b_{t-1}),\ 0 < \alpha < 1, \quad (4)$$

$$\text{Trend: } b_t = \beta(m_t - m_{t-1}) + (1-\beta)b_{t-1},\ 0 < \beta < 1, \quad (5)$$

$$\text{Seasonality: } c_t = \gamma \frac{y_t}{m_t} + (1-\gamma)c_{t-s},\ 0 < \gamma < 1, \quad (6)$$

with $s = 7$ for weekly pattern adjustment. The k-step ahead forecast is then:

$$\hat{y}_{n+k|n} = (m_n + kb_n)(c_{n-s+k}). \quad (7)$$

This last quantity may be used for either smoothing or estimation; i.e., it may be used to replace either the test quantity x_t, as in EWMA_SAGES, or the sliding baseline mean in Eq. 1.[18] From trying both options on SAGES data, GS_SAGES replaces x_t with the Eq. 4 forecast. The standard error in Eq. 1 is stratified by day of week when a weekly pattern is present. Implementation of these equations requires a careful choice of smoothing constants α, β, γ and of initial values $c_0, c_1, \ldots, c_s$, m_0, b_0.

Data Sets for Alerting Algorithm Evaluation

The alerting algorithms were evaluated and compared in two phases. The first phase used authentic reports of dengue/dengue-like illness from the National Epidemiology Center of the Republic of the Philippines (RP). The second phase used simulated, stochastic time-series counts to test algorithm performance on less disease-specific time series with systematic background behavior such as day-of-week effects and temporal correlation.

Authentic Dengue-Related Report Time Series

Collaboration with the RP, the source of the authentic data described in this article, has broadly informed SAGES development. Dengue is a viral infection for which there is no approved vaccine. It remains a severe health threat in the RP and is included in that country's list of officially reportable diseases. Globally, dengue surveillance is important because there are up to 100 million annual infections in tropical climates and because prompt treatment can prevent severe illness.[19]

The patient report data were provided by the National Epidemiology Center of the RP Department of Health. Before 2008, dengue-related illness was reported through the National Epidemic Sentinel Surveillance System (NESSS), which gathered data from multiple hospital surveillance sites. After the 2002 emergence of severe acute respiratory syndrome, surveillance methods changed, and NESSS was replaced by the Philippines Integrated Disease Surveillance and Response system (PIDSR), which became operational in 2008. The number of reportable diseases was increased, and disease-reporting units were added. Some of the data collection, processing tools, and methods were also changed under this new system, enabling access to more complete information.

The time series used for the algorithm evaluation and comparison were regional daily counts of dengue/dengue-like illness reports from the NESSS or PIDSR system covering nearly 19 years, from the beginning of 1993 to the end of 2011. Regions used for aggrega-

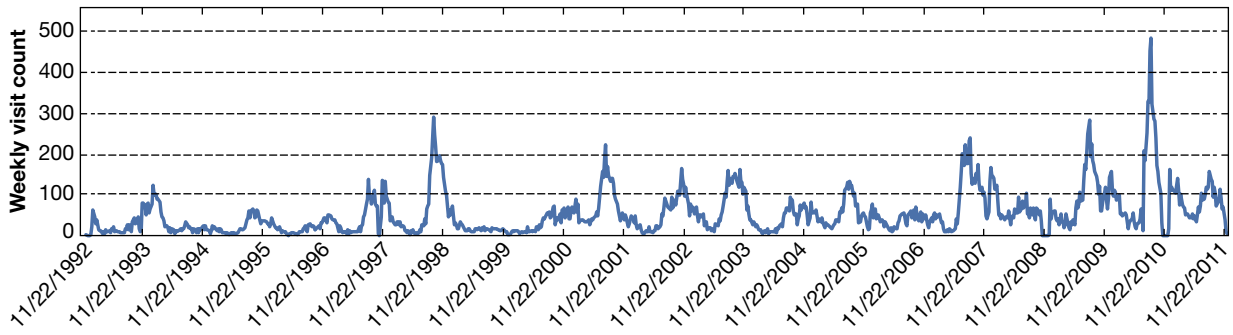

Figure 1. Weekly counts of dengue-related clinic hospital reports in Cebu Province from November 1992 through November 2011.

tion were the entire province of Cebu, Cebu City, and the eight largest municipalities in the province, for a total of 10 time series of report counts. Algorithms were applied to both daily and aggregated weekly time series. Province-level weekly report counts are plotted in Fig. 1. The time series showed no characteristic day-of-week effects or other cyclic or systematic background behavior other than an annual epidemic whose severity, timing, and duration vary from year to year. Compared with less disease-specific time series with complex background features, these data present less of a challenge for anomaly detection, but they are of primary importance because they represent a known annual threat that calls for prompt public health response.

Target Signals for Authentic Dengue Report Count Series

Evaluating the alerting performance of statistical methods requires a sufficient number of target signals in the data. Detailed specification of outbreak dates was unavailable for the 10 PIDSR time series, although dengue epidemics were evident in each time series. These events are associated with the rainy season in RP between June and February, when the mosquito vector for the dengue virus is most prolific.

A total of 96 outbreak intervals were selected from the 10 time series, with at least five intervals from each municipality. The selections were based on experience recognizing aberrant signals in time series from multiple surveillance settings. Although these signal selections are subject to judgment bias, the lack of precise truth data, an obstacle characteristic of biosurveillance research and practice, has led to similar target specification procedures in other method evaluations.[20] Moreover, the procedure in the current effort had nearly 19 years of usable authentic dengue-like case reports from multiple municipalities, clearly visible target events, and a relatively quiet background outside the event intervals.

To avoid the subjective choice of exact beginning and ending dates for an epidemic in the noisier daily data, event dates were chosen from weekly plots and used to evaluate alerting on both daily and weekly data. This decision acknowledged the imprecision of measuring alerting timeliness on authentic daily time series. Alerts during the 96 chosen event intervals were considered true positives, and alerts outside these intervals were considered false positives.

Simulated Syndromic Background Data

The second evaluation phase used time series that were richer, in scale and in background behavior, and more syndromic (i.e., less disease specific). This phase was motivated by the syndromic report-count data increasingly available on a daily basis in resource-limited settings. For example, the SAGES system also collects daily illness data using short message service (SMS) cellular telephone technology,[1] but the SMS data available for the current analysis covered only one full season and were not used in the method evaluation. Lacking sufficient SAGES data history for more syndromic data, we created background data and target signals with simulation.

The simulated time series were modified random vectors drawn from Poisson distributions, often representative of count data. Based on exploratory analysis of the available syndromic SAGES data, we applied three types of modification to these random vectors.

1. **Scale:** Time series were computed for daily Poisson mean (equivalently, variance) levels 0.5, 3, 10, and 50. This range of values gave report count scales from the small municipality level to the province level.
2. **Autocorrelation coefficient:** For each scale level, time series were generated with lag-1 autocorrelation coefficient 0 (for independent daily counts), 0.3, and 0.6 (for extreme next-day dependence). The transformations adding these temporal dependencies preserved the Poisson property.
3. **Day-of-week pattern:** Three day-of-week patterns were applied for each combination of scale and lag-1 autocorrelation. The first pattern was uniform, with no day-of-week weighting, similar to the PIDSR data series. The second pattern was drawn from past syndromic data, with relative weighting of (0.15, 2.15, 1.55, 0.95, 0.95, 0.95, 0.30) for Sunday through Sat-

urday. The third pattern, with weighting vector (0, 2.15, 1.55, 0.95, 0.95, 0.95, 0), assumed no counts on weekends, representing report counts from clinics open only weekdays, a scenario of practical interest in many resource-limited areas.

The modifications described above give 36 combinations of scale, lag-1 autocorrelation, and day-of week effect. For each such combination, 18 stochastic time series of report counts of length 730 days, exactly 2 years, were generated as benchmark data for alerting algorithm evaluation.

Figure 2 gives examples of the simulated daily counts used for algorithm evaluation. Sample simulated series with daily means of 3 and 10 report counts are plotted. The upper half of the figure shows simulated series with no day-of-week effect, whereas series in the lower half are modified with day-of-week proportions extrapolated from authentic report data. As discussed above, 18 such series were generated for each combination of daily mean, autocorrelation coefficient, and day-of-week effect.

Target Signals for Aberration Detection

Target Signals for Simulated Syndromic Count Series

The strategy for injecting target signals into the simulated background data was to select a realistic signal shape, sample from that shape on selected event dates to obtain daily case count injects attributable to an outbreak, and add those injects to the simulated background counts. With this procedure, the signal start, peak, and end dates are known precisely. The strength of the signal was chosen to give a detectable target that would be a challenge to the candidate alerting algorithms. One such injected signal was thus added to each simulated time series. In testing the algorithms, alerts during inject intervals were considered true positives, and alerts on other days false positives. For false alarm calculation, alerts on non-inject days were considered false positives. The advantage of evaluation using simulated data is that no effects of unknown outbreaks are hidden in the data, so that alerts outside of known signal intervals may be accurately called false alerts rather than background alerts.

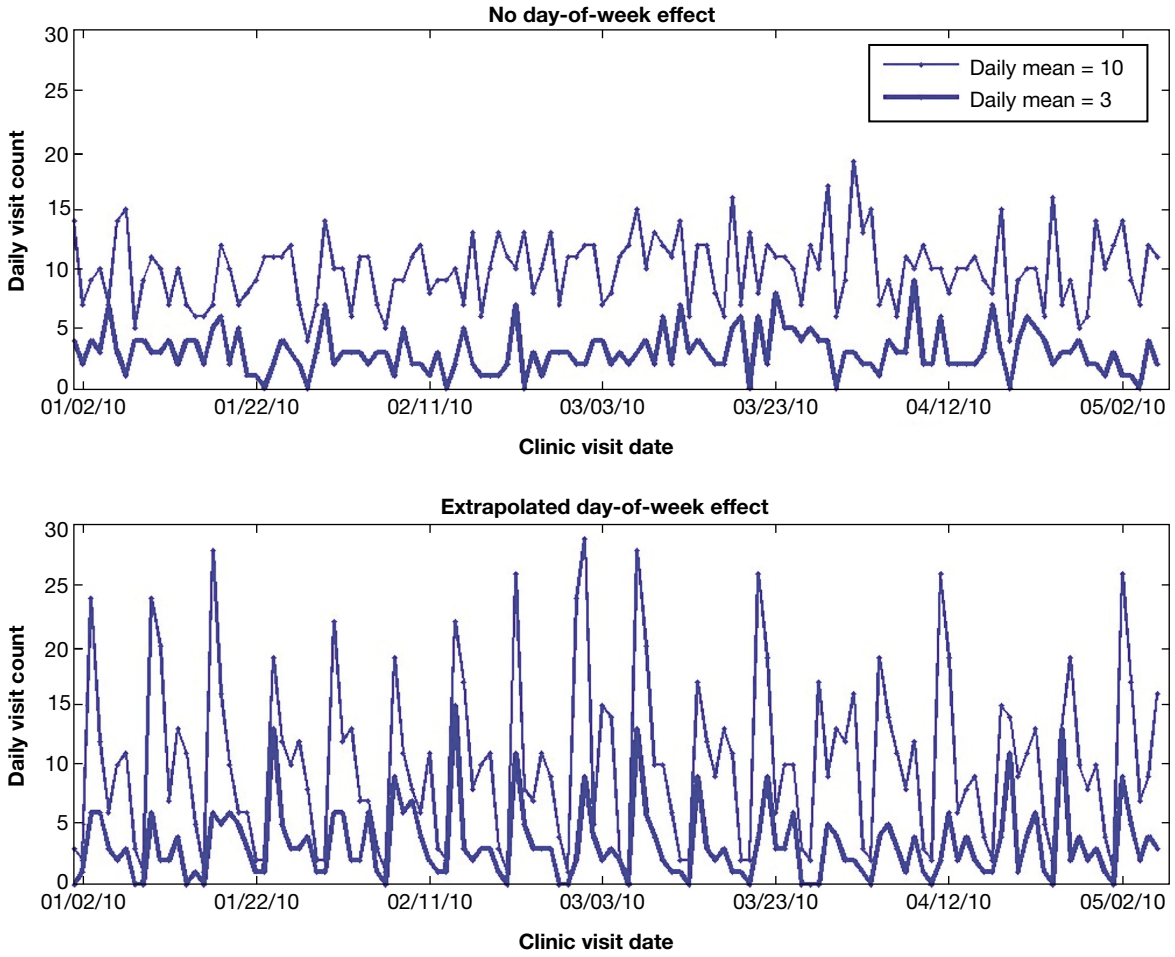

Figure 2. Sample plots of simulated time series without temporal correlation and with daily means of 3 and 10 report counts. Series in the upper plot assume no day-of-week effect, whereas the lower plot shows series with a day-of-week pattern extrapolated from historical data.

For the signal shape, we chose the lognormal distribution proposed by Sartwell in 1950[21] and widely used since then as representative of the epidemic curve distribution of many infectious disease.[10] The rationale is that the distribution of care-seeking dates plausibly reflects the distribution of symptom-onset dates (i.e., the epidemic curve).

Evaluation Measures

Multiple criteria have been recommended for evaluating surveillance algorithms.[22] The results described below apply a combination of these measures including sensitivity, specificity, timeliness, and positive predictive value (PPV). After reviewing published attempts to combine these measures,[23] we sought a straightforward comparison approach emphasizing priorities of the resource-limited setting. We first list operational definitions of the performance measures applied to the authentic dengue data with known events and to the simulated syndromic data with injected events:

- **Event Sensitivity**, the ratio of alerted target events to all target events: Of the total number of events known or injected, for how many does the algorithm alert before the peak date of the event?
- **Specificity**, or (1 – the background alert rate): Of the dates that are not during known or injected events, for what percentage does the algorithm correctly fail to alert?
- **Timeliness:** What is the delay in days between the start of a target event signal and the first algorithm alert?
- **Temporal Sensitivity (Coherence):** On what proportion of the days during target events does the algorithm alert?
- **PPV:** What proportion of all algorithm alerts occur during known or injected events?

Event sensitivity is the value adopted by many authors for surveillance alerting methods and is often weighed against the specificity measure, computed on the basis of event days, not events. The temporal sensitivity or coherence measure is rarely published because during the course of an event lasting more than a week, the data may be anomalous only on a few days around the peak, so values of coherence are typically low, especially in nonspecific data. In practice, however, health monitors often cope with the lack of time and resources by investigating only after several alerts are seen, so the coherence measure was included in this evaluation. The PPV as defined above tells how many alerts are likely to be investigated before an event of interest is found. In practice, this measure is more relevant than specificity, a function of disease prevalence, for evaluating algorithm utility.

From these considerations, we defined an algorithm as an alerting method with a fixed set of parameters and threshold. Recommended algorithms were chosen as follows: Consider only algorithms whose sensitivity, specificity, and coherence meet strict minimum criteria. To accommodate the imprecision in alerting timeliness, drop algorithms whose average alerting timeliness is more than a full day longer than the shortest delay among those remaining. Among the remaining algorithms within 0.005 of the highest PPV value, the one with the shortest average alerting delay was recommended.

RESULTS

Methods tested on the dengue data were Z-score_SAGES, CuSUM_SAGES, and EWMA_SAGES. For the noisier simulated data, typical of less specific syndromic series, the GS_SAGES method was also tested.

Results Using Weekly Dengue-Related Report Data

The testing procedure described in *Methods* was applied to the weekly aggregated counts of the 10 report time series described above to examine alerting performance on the 96 target events. Each series consisted of 992 weeks of counts, with the first 42 weeks reserved as warm-up intervals for the longest algorithm baselines, so that alerts considered false positives could occur in 950 weeks minus the event intervals.

Tested parameter/threshold sets included 120 parameter/threshold combinations for the Z-score_SAGES method, 720 combinations for CuSUM_SAGES to test values of k, and 840 combinations for EWMA_SAGES, including additional combinations for the latter two methods to vary the k and ω constants. These 1680 algorithm combinations were applied to test detection performance on the 96 target signals. From the resulting algorithm output, candidate methods were restricted to those with specificity > 95%, specificity > 95%, coherence > 66%, and alerting delays less than 1.5 weeks. The remaining combinations are tabulated in Table 1 with sorted PPV values above 0.70. No Z-score combination met the timeliness criterion with high PPV. Detection performance for the tabulated combinations is favorable because the weekly data aggregation yields good coherence and high PPV at the required sensitivity/specificity levels for the seasonal target events. Many other CuSUM and EWMA combinations yielded slightly lower PPV values. Analysis of the full table indicated the use of CuSUM_SAGES, and the parameter combination in the second row with an alerting threshold of $p = 0.01$ can be considered the best PPV/timeliness result.

Results Using Daily Dengue-Related Report Data

The same methods were applied with a similar number of parametric combinations to the 10 daily dengue-report time series without aggregation. Each series contained 6939 days of counts, with the first 300 days

reserved for algorithms with long baselines, so that false positives could occur in each series on 6639 days minus the event intervals. Because of the increased volatility of daily count data, few algorithm combinations yielded a coherence measure above 50%. From the resulting algorithm output, candidate methods were restricted to those with sensitivity > 95%, specificity > 95%, coherence > 40%, and alerting delays less than 4.5 days. The remaining CuSUM_SAGES and EWMA_SAGES combinations are tabulated in Table 2. Again, no Z-score_SAGES combination met the timeliness criterion with acceptable PPV.

From Table 2, 16 algorithm combinations met the event sensitivity/specificity requirement with PPV values above 0.7.

Coherence values are lower and alerting thresholds higher than in Table 1. The EWMA_SAGES algorithm in the top row achieved a PPV of 0.77 with a mean delay

Table 1. Algorithms/parameter/threshold combinations yielding best alerting performance for weekly report count data

Algorithm	Baseline Length (weeks)	Guard Band Length (weeks)	Parameter (CuSUM k, EWMA ω)	Alerting Threshold	Specificity	Event Sensitivity	Coherence	PPV	Delay (weeks)	Delay (days)
CuSUM_SAGES	12	2	0.3	0.01	0.96	0.99	0.73	0.75	1.48	10.39
CuSUM_SAGES	12	2	0.2	0.01	0.96	0.99	0.75	0.75	1.39	9.74
CuSUM_SAGES	12	2	0.4	0.02	0.96	1.00	0.78	0.75	1.42	9.96
CuSUM_SAGES	12	2	0.5	0.03	0.95	1.00	0.79	0.74	1.45	10.18
CuSUM_SAGES	12	2	0.1	0.01	0.95	1.00	0.77	0.73	1.35	9.45
CuSUM_SAGES	12	1	0.4	0.02	0.96	1.00	0.74	0.73	1.46	10.25
CuSUM_SAGES	12	2	0.07	0.01	0.95	1.00	0.77	0.73	1.37	9.60
CuSUM_SAGES	12	1	0.5	0.03	0.95	1.00	0.76	0.72	1.48	10.39
EWMA_SAGES	12	2	0.7	0.03	0.96	1.00	0.71	0.71	1.49	10.46

Table 2. Algorithms/parameter/threshold combinations yielding best alerting performance for daily report count data

Algorithm	Baseline Length (days)	Guard Band Length (days)	Parameter (CuSUM k, EWMA ω)	Alerting Threshold	Specificity	Event Sensitivity	Coherence	PPV	Delay (days)
EWMA_SAGES	112	7	0.5	0.05	0.97	0.98	0.45	0.77	3.95
CuSUM_SAGES	112	7	0.3	0.05	0.97	0.99	0.49	0.77	4.05
CuSUM_SAGES	112	7	0.2	0.05	0.97	0.99	0.56	0.76	4.11
CuSUM_SAGES	112	7	0.4	0.05	0.97	0.99	0.46	0.76	3.69
CuSUM_SAGES	112	2	0.3	0.05	0.97	0.99	0.46	0.76	3.96
CuSUM_SAGES	112	2	0.2	0.05	0.97	0.99	0.53	0.76	3.95
CuSUM_SAGES	112	7	0.5	0.05	0.97	1.00	0.42	0.75	3.72
CuSUM_SAGES	112	2	0.4	0.05	0.97	0.99	0.43	0.75	3.70
CuSUM_SAGES	112	7	0.1	0.05	0.96	0.97	0.66	0.75	4.27
CuSUM_SAGES	112	2	0.1	0.05	0.96	0.97	0.63	0.75	4.30
EWMA_SAGES	84	7	0.5	0.05	0.97	0.99	0.41	0.73	4.21
EWMA_SAGES	112	7	0.3	0.05	0.96	1.00	0.52	0.72	3.38
CuSUM_SAGES	84	7	0.3	0.05	0.97	0.99	0.44	0.71	4.10
CuSUM_SAGES	84	7	0.4	0.05	0.97	0.99	0.41	0.70	4.01
EWMA_SAGES	112	2	0.3	0.05	0.96	1.00	0.48	0.70	3.53
CuSUM_SAGES	84	7	0.2	0.05	0.96	0.99	0.51	0.70	4.18

less than 4 days. However, a user that values repeated alerting during an event (higher coherence) could prefer the CuSUM_SAGES combinations with k-values below 0.3.

Results Using Simulated Syndromic Data

Relative Method Performance

Compared with alerting performance on the PIDSR dengue report data, the algorithms were less effective on the simulated syndromic time series. Coherence and positive predictive value dropped consistently because of the higher background noise level of Fig. 1, day-of-week effects for some of the series, and the transient, stochastic nature of the target signals.

For each combination of simulation parameters, the alerting method/parameter combination was the one with the highest PPV given specificity and event sensitivity ≥ 95%, coherence > 20%, and alerting timeliness within 1 day of other qualifying combinations. The top methods by these criteria are listed in Table 3 with the corresponding coherence, PPV, and alerting delay. The GS_SAGES method was the choice for series with a day-of-week effect whenever the count mean was at least 3, likely because of the GS_SAGES adaptation to cyclic patterns. For the sparse time series with mean value 0.5, a CuSUM or EWMA was the best choice. As in the results from the less noisy dengue data with clearer seasonal target events, none of the Z-score_SAGES algorithm combinations met the combined criteria.

Table 3. Optimal alerting method and values of coherence, PPV, and alerting delay for simulated daily report count time series with each combination of report count mean, lag-1 autocorrelation coefficient, and day-of-week effect

Mean Daily Visit Count	Temporal Correlation Coefficient (lag-1)	Day-of-Week Effect	Alerting Method	Coherence	PPV	Alerting Delay (days)
50	0	No	EWMA	0.33	0.58	3.89
50	0.3	No	CuSUM	0.33	0.41	4.47
50	0.6	No	GS_SAGES	0.25	0.37	4.32
50	0	Yes	GS_SAGES	0.26	0.41	4.26
50	0.3	Yes	GS_SAGES	0.28	0.33	4.63
50	0.6	Yes	GS_SAGES	0.20	0.32	4.32
10	0	No	CuSUM/EWMA	0.24	0.52	3.89
10	0.3	No	GS_SAGES	0.23	0.32	4.79
10	0.6	No	GS_SAGES	0.29	0.19	3.21
10	0	Yes	GS_SAGES	0.35	0.36	2.79
10	0.3	Yes	GS_SAGES	0.22	0.37	4.05
10	0.6	Yes	GS_SAGES	0.23	0.22	4.11
3	0	No	EWMA	0.38	0.31	3.89
3	0.3	No	EWMA	0.40	0.22	3.42
3	0.6	No	GS_SAGES	0.32	0.12	3.05
3	0	Yes	GS_SAGES	0.28	0.25	3.05
3	0.3	Yes	GS_SAGES	0.23	0.15	3.58
3	0.6	Yes	GS_SAGES	0.29	0.20	3.05
0.5	0	No	CuSUM/EWMA	0.38	0.16	2.42
0.5	0.3	No	EWMA	0.26	0.11	3.26
0.5	0.6	No	EWMA	0.36	0.10	3.11
0.5	0	Yes	CuSUM/EWMA	0.31	0.14	3.58
0.5	0.3	Yes	CuSUM	0.23	0.11	3.84
0.5	0.6	Yes	EWMA	0.30	0.12	3.58

Overall Alerting Quality

The effects of the data scale and of autocorrelation on PPV are summarized in the bar charts in Fig. 3 for series with and without day-of-week effects. The improvement of PPV with the scale of the data is consistent. For sparse data series, PPV is near or below 0.1, indicating that only one of 10 alerts results from an injected event. For means of 10 or 50, the PPV exceeds 0.3 unless autocorrelation is excessive.

Algorithm Recommendations

This section summarizes method recommendations depending on the amount of historical data available, the data scale represented by the series mean, and the day-of-week effect. The chart in Table 4 suggests the method of choice given the available data. The baseline columns represent 2-, 4-, 8-, and 16-week baseline lengths. If 112 days of representative data are available, the methods in the rightmost column are recommended.

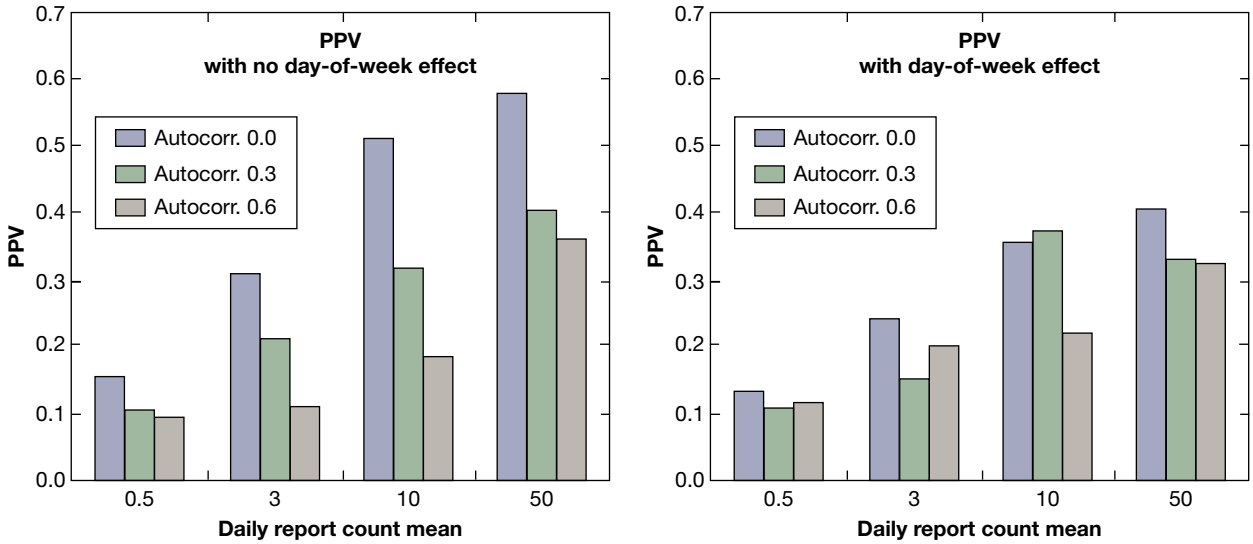

Figure 3. Bar chart comparisons of best PPV achieved using simulated time series as a function of daily report count mean, lag-1 autocorrelation coefficient, and presence of day-of-week effect.

For PPV averaged over all algorithm/parameter combinations, the overall dependence of PPV on the baseline length and time series scale is summarized in Fig. 4. Except for sparse data series, a baseline of at least 28 days is recommended in view of the jump in PPV from 14 to 28 days. Additional PPV increases may be realized when longer data history of up to 56 days is reliably available. Previous experience and other studies[10] have suggested diminishing returns for baselines longer than 8 weeks except for regression methods developed for specific time series with mean values above 10 encounters per day.

CONCLUSIONS

The above-described algorithm evaluation analyses on authentic and simulated data support the use of alerting methods on surveillance data from resource-limited settings. The analyses give the background and usage guidance for the open-source methods provided with SAGES.[7] In view of the wide range of backgrounds and experience of SAGES users, only basic knowledge of the local data characteristics—e.g., outcomes of interest and the scale, seasonal/cyclic behavior, and quality of indicator time series—is required to use this guidance.

From the analysis on the dengue report data, if input series counts are aggregated from carefully chosen patient records, such as the Cebu dengue-related patient reports, then adaptive statistical methods can yield timely alerts with high sensitivity and specificity and practical PPV. The weekly analysis gave better overall PPV and coherence statistics at the cost of alerting timeliness. From the best timeliness results using the weekly data, alerts are not expected until the second week of an event. By contrast, analyses using daily data consistently showed that alerts could be expected after 4–5 days. However, this timeliness advantage has relevance only

Table 4. Chart of recommended alerting methods for daily report count data as a function of baseline length, daily mean count, and presence of day-of-week effect

Recommended Alerting Methods	Daily Mean Visits	Baseline length (days)			
		14	28	56	112
No Day-of-Week Effect	0.5	CuSUM, EWMA	EWMA	EWMA	CuSUM, EWMA
	3	CuSUM, EWMA	CuSUM, EWMA	CuSUM, EWMA	CuSUM, EWMA
	10	CuSUM	CuSUM	CuSUM, EWMA	EWMA
	50	CuSUM	CuSUM, EWMA	CuSUM, EWMA	EWMA
Customary Day-of-Week Effect	0.5	CuSUM	EWMA	EWMA	EWMA
	3	CuSUM, EWMA	CuSUM, EWMA	GS_SAGES	GS_SAGES
	10	CuSUM, EWMA	GS_SAGES	GS_SAGES	GS_SAGES
	50	GS_SAGES	GS_SAGES	GS_SAGES	GS_SAGES

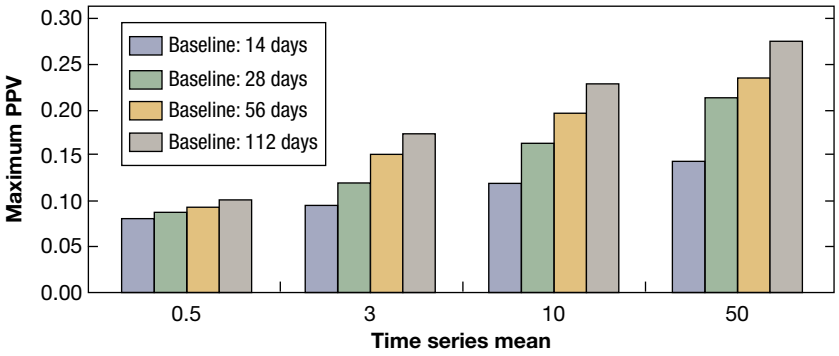

Figure 4. Dependence of alerting algorithm PPV on algorithm baseline length and time series scale.

if the public health system can initiate response measures soon enough to exploit it. There will always be the question of whether an early alert is a true signal, and the reduced coherence of algorithms applied to the more volatile daily data will require confidence in the analytic methods and probably the need for additional corroboration. Thus, operational considerations and investigation protocols should be considered along with statistical alerting capability in designing a surveillance system.[24] The analyses on the simulated syndromic data show that alerting performance can vary widely with the input data, emphasizing the importance of careful selection of data indicators or syndrome groups and of monitoring only as many as the investigation and response resources can manage, as noted by other authors.[25,26] For many nonspecific syndrome groups, the PPV of even a well-chosen method may be much lower than 0.2–0.3, and correspondingly many alerts will need to be investigated before an event of interest is detected. Especially in resource-limited surveillance, this requirement must be understood, or the alerts will be ignored.

Regarding the methods themselves, the analyses show that effective monitoring requires methods appropriate for and adaptive to the input data. Results with the Z-score_SAGES method show that the standard Shewhart-type control chart is not suitable for surveillance data streams, which typically violate the underlying data assumptions. Adaptive versions of this method would be more competitive for detection of single spikes, whereas the other methods are better suited to detection of signals spread over multiple days. Depending on the record filtering criteria, input data streams may display distinct day-of-week patterns. Only the GS_SAGES method yielded effective performance measures while controlling for these patterns. In the absence of these patterns, certain parameter combinations of both EWMA and CuSUM methods were optimal. None of the methods gave high PPV for alerting of moderate-sized events in sparse data streams.

There are limitations to evaluations using both authentic and the simulated data streams. Authentic historical background data pose the problem that false alerts cannot be verified, so calculated PPV may be underestimated. Authentic target signals, the footprints of health events in data streams, are rarely available, and the beginning and ending dates of these signals must be estimated, as in the dengue report data described above. This uncertainty compromises the alerting timeliness and coherence measures. The Cebu dengue report counts were chosen to lessen these problems. Simulated data streams and target signals avoid these problems, but evaluations using simulated data have the burden of proving that the results are applicable to authentic data. The simulated time series described above were designed to capture statistical properties of observed patient record counts. The combination of authentic and simulated data testing was intended to supply evidence from both perspectives.

Future efforts will seek to further customize statistical alerting methods for surveillance data streams, but the direction of such improvements must keep pace with developing needs in response to global epidemiological needs and with advancing data technology.

ACKNOWLEDGMENTS: The authors acknowledge the support and advice of Erhan Guven and rest of the APL SAGES technical team under Program Manager Sheri Lewis. The opinions or assertions in this article are the private views of the authors and do not necessarily reflect the official policy or position of the U.S. Department of the Army, the U.S. Department of Defense, or the U.S. government.

REFERENCES

[1]Lewis, S. L., Feighner, B. H., Loschen, W. A., Wojcik, R. A., Skora, J. F., et al., "SAGES: A Suite of Freely-Available Software Tools for Electronic Disease Surveillance in Resource-Limited Settings," *PLoS One* **6**(5), e19750 (2011).

[2]Fifty-Eighth World Health Assembly, "Revision of the International Health Regulations," WHA58.3, http://www.who.int/csr/ihr/IHRWHA58_3-en.pdf (23 May 2005).

[3]Homeland Security Presidential Directive 21 (HSPD-21), "Public Health and Medical Preparedness," https://www.fas.org/irp/offdocs/nspd/hspd-21.htm (18 Oct 2007).

[4]Chretien, J. P., Burkom, H. S., Sedyaningsih, E. R., Larasati, R. P., Lescano, A. G., et al., "Syndromic Surveillance: Adapting Innovations to Developing Settings," *PLoS Med.* **5**(3), e72 (2009)

[5]May, L., Chretien, J.-P., and Pavlin, J. A., "Beyond Traditional Surveillance: Applying Syndromic Surveillance to Developing Settings—Opportunities and Challenges," *BMC Public Health* **9**(242), 1–11 (2009).

[6]Lombardo, J., Burkom, H., Elbert, E., Magruder, S., Happel-Lewis, S., et al., "A System Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II)," *J. Urban Health* **80**(2 Suppl 1), i32–i42 (2003).

[7]APL, "Tools," SAGES website, http://www.jhuapl.edu/Sages/tools.html (accessed 2 Jan 2014).

[8]Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., and Andrews, N., "Statistical Methods for the Prospective Detection of Infectious Disease Outbreaks: A Review," *J. R. Statist. Soc. A* **175**(1), 49–82 (2012).
[9]Ryan, T. P., *Statistical Methods for Quality Improvement*, John Wiley & Sons, New York (1989).
[10]Burkom, H. S., "Development, Adaptation, and Assessment of Alerting Algorithms for Biosurveillance," *Johns Hopkins APL Tech. Dig.* **24**(4), 335–342 (2003).
[11]Hutwagner, L., Thompson, W. W., Seeman, G. M., and Treadwell, T., "A Simulation Model for Assessing Aberration Detection Methods in Public Health Surveillance for Systems with Limited Baselines," *Statist. Med.* **24**(4), 543–550 (2005).
[12]Lombardo, J. S., and Ross, D. "Disease Surveillance: A Public Health Priority," *Disease Surveillance: A Public Health Informatics Approach*, J. S. Lombardo and D. L. Buckeridge (eds.), John Wiley & Sons, Inc., Hoboken, pp. 1–39 (2007).
[13]Hawkins, D. M., and Olwell, D. H., *Cumulative Sum Control Charts and Charting for Quality Improvement*, Springer, New York (1998).
[14]Fricker, R. D., Hegler, B., and Dunfee, D. A., "Comparing syndromic Surveillance Detection Methods: EARS' versus a CUSUM-Based Methodology," *Statist. Med.* **27**(17), 3407–3429 (2008).
[15]Lucas, J. M., and Crosier, R. B., "Fast Initial Response for CUSUM Quality Control Schemes: Give Your CUSUM a Head Start," *Technometrics* **24**(3), 199–205 (1982).
[16]Elbert, Y., and Burkom, H., "Development and Evaluation of a Data-Adaptive Alerting Algorithm for Univariate Temporal Biosurveillance Data," *Statist. Med.* **28**(26), 3226–3248 (2009).
[17]Burkom, H., Murphy, S. P., and Shmueli G., "Automated Time Series Forecasting for Biosurveillance," *Statist. Med.* **26**(22), 4202–4218 (2007).
[18]Stoto, M. A., Schonlau, M., and Mariano, L. T., "Syndromic Surveillance: Is it Worth the Effort?" *Chance* **17**(1), 19–24 (2004).
[19]World Health Organization, "Dengue and Severe Dengue," Fact Sheet no. 117, http://www.who.int/mediacentre/factsheets/fs117/en/ (updated Sep 2013).
[20]Siegrist, D., and Pavlin, J., "Bio-ALIRT Biosurveillance Detection Algorithm Evaluation," *MMWR Morb. Mortal. Wkly. Rep.* **53**(Suppl), 152–158 (2004).
[21]Sartwell, P. E., "The Distribution of Incubation Periods of infectious Disease," *Am. J. Hyg.* **51**, 310–318 (1950).
[22]Buehler, J. W., Hopkins, R. S., Overhage, J. M., Sosin, D. M., and Tong. V., "Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks: Recommendations from the CDC Working Group," *MMWR Recomm. Rep.* **53**(RR05), 1–11 (2004).
[23]Kleinman, K. P., and Abrams, A. M., "Assessing Surveillance Using Sensitivity, Specificity and Timeliness," *Stat. Methods Med. Res.* **15**(5), 445–464 (2006).
[24]Soto, G., Araujo-Castillo, R. V., Neyra, J., Fernandez, M., Leturia, C., et al., "Challenges in the Implementation of an Electronic Surveillance System in a Resource-Limited Setting: Alerta, in Peru" *BMC Proc.* **2**(Suppl 3):S4 (2008).
[25]Lescano, A. G., Larasati, R. P., Sedyaningsih, E. R., Bounlu, K., Araujo-Castillo, R. V., et al., "Statistical Analyses in Disease Surveillance Systems," *BMC Proc.* **2**(Suppl 3): S7 (2008).
[26]Venkatarao, E., Patil, R. R., Prasad, D., Anasuya, A., and Samuel, R., "Monitoring Data Quality in Syndromic Surveillance: Learnings from a Resource Limited Setting," *J. Glob. Infect. Dis.* **4**(2), 120–127 (2012).

The Authors

Howard S. Burkom is an APL Principal Professional Staff member who leads and designs algorithm development initiatives for the ESSENCE and SAGES systems. **Yevgeniy A. Elbert** is an APL statistician who contributes to all phases of data analysis and method implementation and evaluation. **Jacqueline S. Coberly** is an APL epidemiologist who applies her academic and field experience to ensure relevance and practicality of technical development, to participate in evaluation, and to advise/assist with documentation. Three medical epidemiologists from the Republic of Philippines (RP) were essential to this study, not only for arranging access to critical data sets but also for providing perspective on the challenges of doing surveillance in the RP and for description of national surveillance programs. They are **John Mark Velasco** of the Armed Forces Research Institute of Medical Sciences and **Enrique A. Tayag** and **Vito G. Roque Jr**. of the National Epidemiology Center at the Department of Health. For further information on the work reported here, contact Howard Burkom. His e-mail address is howard.burkom@jhuapl.edu.

The *Johns Hopkins APL Technical Digest* can be accessed electronically at **www.jhuapl.edu/techdigest**.