

# Approaches to Syndromic Surveillance When Data Consist of Small Regional Counts

Peter A. Rogerson, I. Yamada  
University at Buffalo, Buffalo, New York

**Corresponding author:** Peter A. Rogerson, University at Buffalo, Department of Geography, Wilkeson Hall, Buffalo, NY 14261. Telephone: 716-645-2722; Fax: 716-645-2329; E-mail: rogerson@buffalo.edu.

## Abstract

**Introduction:** Statistical systems designed for syndromic surveillance often must be able to monitor data received simultaneously from multiple regions. Such data might be of limited size, which would eliminate the possibility of using more common surveillance methods that assume data from a normal distribution.

**Objectives:** The objectives of this study were to design and illustrate a multiregional surveillance system based on data inputs consisting of small regional counts, where frequencies are typically on the order of  $\leq 5$ .

**Methods:** Cumulative sum (CUSUM) methods designed for cumulating the sum of the deviations between observed and expected Poisson-distributed data were modified to account for changing expectations over time, including weekly and monthly effects. Data on lower respiratory tract infections during 1996–1999 at multiple Boston clinics among residents from 287 census tracts were used to illustrate the approach.

**Results:** When each region was monitored, 19% of the census tracts signaled a departure during 1999 from the base period (1996–1998) rates. When local statistics were used to monitor tracts and neighborhoods consisting of surrounding tracts, 60% of tracts experienced departures during 1999 from the base period. These results imply that the increases in lower respiratory tract infection that occurred during 1999 were geographically pervasive.

**Conclusions:** Poisson CUSUM methods are useful for monitoring small regional counts over time. The methods can be generalized to account for time-varying expectations in the counts.

## Introduction

Detecting the locations of statistically significant increases in the rates of health syndromes among multiple geographic areas as rapidly as possible is a critical public health need (1). Multiple systems are being designed to achieve this goal; comprehensive discussion of the desirable features of a statistical health surveillance system has been published previously (2). This paper focuses on two characteristics of such systems: 1) systems should be capable of detecting increases in regional rates quickly while keeping the number of false alerts at an acceptable level, and 2) observations might consist of limited frequencies that would necessitate the use of binomial or Poisson variables instead of normally distributed variables.

Multiple approaches to spatial surveillance in a public health context have been taken previously. One approach is to use cumulative sum (CUSUM) methods to monitor disease counts in geographic areas of interest (3). Another is to perform surveillance by detecting outliers in a temporal sequence of observed binomial variables for multiple geographic regions (4). Other investigators take existing spatial statistical methods used for retrospective detection of geographic clusters of disease and modify them for use in surveillance, which requires repeated tests for emergent clusters (5–7).

This paper uses and develops further a CUSUM approach for small counts (i.e., where frequencies are typically on the order of  $\leq 5$ ) assumed to follow a Poisson distribution. CUSUM methods cumulate deviations between observed and expected counts during a given period and generate an alert or signal when cumulated observed counts exceed expected counts by a predetermined threshold (8).

This paper reviews CUSUM methods for normal and Poisson-distributed variables. It then describes how to modify the Poisson CUSUM approach to allow the expected counts to vary from one period to the next. It also indicates how the approach can be used to monitor neighborhoods consisting of a set of contiguous regional units. These approaches are applied to data on lower respiratory infection episodes reported by Boston-area clinicians during January 1996–October 1999. The paper concludes with a discussion of findings.

## CUSUM Methods

CUSUM methods are designed to detect sudden changes in the mean value of a quantity of interest; they are widely used in industrial process control to monitor production quality. The basic methods rely on two assumptions: 1) the quan-

tity being monitored is distributed normally, and 2) the variable exhibits no serial autocorrelation.

If the variable of interest is converted to a  $z$ -score with mean 0 and variance 1, the CUSUM, following observation  $t$ , is defined as follows:

$$S_t = \max(0, S_{t-1} + z - k)$$

where  $k$  is a parameter. A change in mean is signaled if  $S_t > h$ , where  $h$  is a threshold parameter.

Values of  $z$  in excess of  $k$  are cumulated. The parameter  $k$  in this instance, in which a standardized variable is being monitored, is often chosen to be equal to one half; in the more general case,  $k$  is often chosen to be equal to one half the standard deviation associated with the variable being monitored.

The parameter  $h$  is chosen in conjunction with a predetermined acceptable rate of false alerts; high values of  $h$  lead to a low probability of a false alert but also a lower probability of detecting a real change. The time between false alerts is the in-control average run length and is designated by the notation  $ARL_0$ . When  $k = 1/2$ , an approximation for  $ARL_0$  is

$$ARL_0 = 2(e^a - a - 1)$$

where  $a = h + 1.166$  (9). One can choose the parameter  $h$  by first deciding upon a value of  $ARL_0$ , and then solving the approximation for the corresponding value of  $h$ . This expression for the average run length can be solved, approximately, for  $h$  (P. Rogerson, University at Buffalo, unpublished data):

$$h \approx \left( \frac{ARL_0 + 4}{ARL_0 + 2} \right) \ln \left( \frac{ARL_0}{2} + 1 \right) - 1.166$$

The choice of  $k = 1/2$  minimizes the time required to detect a 1 standard-deviation increase in the mean. More generally,  $k$  is chosen to be equal to one half the size of the change (in units of standard deviations) sought for rapid detection. For this case (i.e., when  $k$  might take on a value other than one half)

$$h \approx \left( \frac{2k^2 ARL_0 + 2}{2k^2 ARL_0 + 1} \right) \frac{\ln(1 + 2k^2 ARL_0)}{2k} - 1.166$$

## CUSUMs for Poisson Variables

When the assumption of normality is not a good one, transformations to normality are sometimes possible. One such normalizing transformation for data consisting of small counts is (10):

$$y = \frac{x - 3\lambda + 2\sqrt{\lambda x}}{2\sqrt{\lambda}}$$

where  $x$  is the observed count and  $\lambda$  is the expected count.

This transformation can be misleading for small values of  $\lambda$ . In particular, the actual  $ARL_0$  values might differ substantially from the desired nominal values. For example, when desired values of  $ARL_0 = 500$  and  $ARL_1 = 3$  (where  $ARL_1$  is the average time taken to detect an increase) are used in situations where  $\lambda < 2$ , simulations demonstrate that using this transformation will almost always yield actual values of  $ARL_0$  substantially lower than the desired value of 500. In certain cases (e.g.,  $\lambda \approx 0.15$ ), the actual ARL will be  $< 100$ , indicating a much higher rate of false alerts than desired. The performance is better when  $ARL_0 = 500$  and  $ARL_1 = 7$ , but use of the transformation will again lead to substantially more false alerts than desired when  $\lambda$  is less than approximately 0.25. Also troubling is the instability with respect to similar values of  $\lambda$ ;  $\lambda = 0.56$  will lead to an  $ARL_0$  of approximately 400, whereas  $\lambda = 0.62$  is associated with an  $ARL_0$  of  $> 700$ . This is also true when  $ARL_1 = 3$ ;  $\lambda = 0.96$  has an ARL of approximately 212, whereas  $\lambda = 0.98$  has an ARL of 635.

When the variable being monitored has a Poisson distribution, the CUSUM is

$$S_t = \max(0, S_{t-1} + X_t - k)$$

New considerations are necessary to determine the parameters  $k$  and  $h$  (12). If  $\lambda_0$  is the mean value of the in-control Poisson parameter, the  $k$ -value that minimizes the time to detect a change from  $\lambda_0$  to a prespecified out-of-control parameter  $\lambda_1$  is

$$k = \frac{\lambda_1 - \lambda_0}{\ln \lambda_1 - \ln \lambda_0} \quad (1)$$

Then,  $h$  can be determined from the values of the parameter  $k$  and the desired  $ARL_0$  by using either a table (11), Monte Carlo simulation, or an algorithm that makes use of a Markov chain approximation (12).

Poisson CUSUM methods have been applied previously in a public health context, primarily in surveillance of congenital malformations (13,14); the approach has also been recommended in surveillance for *Salmonella* outbreaks (15).

## Poisson CUSUM Methods with Time-Varying Expectations

The expected in-control value associated with the Poisson variable might vary with time ( $\lambda_{0,t}$ ;  $t = 1, 2, \dots$ ) (e.g., as a result of seasonal effects). Simply implementing a CUSUM scheme with constant parameters would have misleading results if the actual values of  $\lambda_0$  fluctuated from period to period about the constant assumed parameter. Instead, time-specific values of the parameters  $k$  and  $h$  were used. The observed values,  $X_t$ , were then used in the CUSUM as follows:

$$S_t = \max[(0, S_{t-1} + c_t(X_t - k_t)] \quad (2)$$

where the parameters  $c_t$  and  $k_t$  change from one period to the next, and their values are now discussed.

First  $h$  is chosen on the basis of the mean of the time-varying Poisson parameter, an associated value of  $k$ , and the desired  $ARL_0$ . Once  $h$  is chosen, next choose  $k_t$  on the basis of  $\lambda_{0,t}$  and  $\lambda_{1,t}$ ,

$$k_t = \frac{\lambda_{1,t} - \lambda_{0,t}}{\ln \lambda_{1,t} - \ln \lambda_{0,t}} \quad (3)$$

Then,  $c_t$  is chosen as the ratio  $h$  to  $h_t$ , where the latter is the value of the threshold associated with the desired  $ARL_0$ ,  $k_t$ , and constant values of  $\lambda_{0,t}$  and  $\lambda_{1,t}$ . Thus,  $c_t = h/h_t$ . The quantity  $c_t$  is chosen so that observed counts  $X_t$  will make the proper relative contribution toward the signaling parameter  $h$  that is used in the actual CUSUM. If, for example,  $h > h_t$ , then the contribution  $X_t - k_t$  is scaled up by the factor  $h/h_t$ . An alternative approach is to apply a multiplicative factor to the baseline, or average value of  $\lambda$  (16).

## Poisson CUSUM Methods for Neighborhoods Consisting of Contiguous Regional Units

An extension is to construct local statistics in association with each geographic unit. These are defined as a weighted sum of the region's observation and surrounding observations, where the weights could decline with increasing distance from the region. CUSUMs associated with these local statistics would be monitored. Local statistics are spatially autocorrelated, and Monte Carlo simulation of the null hypothesis can be performed to determine appropriate thresholds for the CUSUMs if no deviation from expected values of the Poisson parameters exists.

## Application to Boston Data on Lower Respiratory Infection

### Data

Harvard Vanguard Medical Associates (Boston, Massachusetts) uses an automated record system for its 14 clinics. After each patient office visit, the clinician records diagnoses and *International Classification of Disease, Ninth Revision* (ICD-9) codes. Patient addresses are recorded; these have been geocoded and assigned to census tracts.

Data on lower respiratory infection episodes were available for January 1996–October 1999. During this period, 47,731 episodes occurred that could be assigned to one of the 287 census tracts in the study region.

### Model for Expected Counts

The first 3 years of data (January 1996–December 1998) were used to calibrate logistic regression models for each census tract. The logistic transform of the probability of a visit is taken to be a linear function of the explanatory variables:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_m x_{mi} + \varepsilon_i$$

where  $p_i$  is the probability of a visit in region  $i$ ;  $x_{li}$  is the value of explanatory variable  $l$  in region  $i$ ; and the  $\beta$ s are the regression coefficients;  $m$  explanatory variables and  $m + 1$  coefficients are estimated in each tract.

Compared with the random-effects model described previously (4), this modeling approach has coefficients that are specific to individual regions. However, constructing a model for each region might result in region-specific coefficients that might not be reliable over time, especially when they are estimated from a limited number of observations. An alternative might be to have region-specific dummy variables in a single equation, but this could use a substantial number of degrees of freedom relative to the number of observations.

In each census tract, the unit of observation was the day. During the 3-year base period (i.e., 1,096 days), expected counts on each day were modeled as a function of time trend (i.e., the logistic transform of the probability of a visit was taken to be a linear function of the day number). Eleven dummy variables were created for the months of the year; December was taken as the arbitrary, omitted category. Finally, a dummy variable was also included for visits that occurred on weekends, with weekday observations as the reference category. Another potential variable capturing temporal autocorrelation in the counts was also considered, but in the majority of cases it was not significant. Inclusion of such a variable would be a way to address violations of the assumption of independence in the CUSUM method.

The average coefficient for each of the explanatory variables (in which the average is taken over the 287 census tracts) is provided (Table 1). Visits are most likely in December; the probability of visits declines steadily thereafter until July. In August, the probability of a visit begins to increase, until reaching its maximum in December. The likelihood of weekend visits is substantially lower than weekday visits, as expected. Finally, the average time trend is positive.

**TABLE 1. Average coefficients in logistic regression model, by month\* and day†**

Variable	Coefficient
January	-0.211
February	-0.472
March	-0.692
April	-0.976
May	-1.113
June	-1.527
July	-2.039
August	-1.419
September	-1.369
October	-0.657
November	-0.270
Weekend	-1.262
Day	0.00345
Intercept	-7.640

\* December is the omitted reference month.

† Refers to the time trend; the coefficient indicates the daily increase in the log-odds of a visit.

## Poisson CUSUM Method

For an illustration of how the modified CUSUM approach might be applied, the estimated parameters for each tract were used, together with the relevant explanatory variables, to derive the expected probability of a visit for each day, for each census tract, for the 303-day period beginning January 1, 1999. These expected probabilities were multiplied by the number of patients in each tract on each day to derive the expected number of visits on each day. The latter quantity is the time-varying, in-control Poisson parameter,  $\lambda_{0,t}$ . To minimize the time to detect a one half standard-deviation change in this parameter, the out-of-control Poisson parameter is chosen to be

$$\lambda_{1,t} = \lambda_{0,t} + \frac{1}{2} \sqrt{\lambda_{0,t}}$$

Although minimizing the time to detecting a one standard-deviation change is probably more common, one half of a standard deviation is used here because the standard deviation is so large relative to the mean. For example, when

$$\lambda_{0,t} = 0.1, \sqrt{\lambda_{0,t}} = 0.32$$

for detecting a 1 standard-deviation change,

$$\lambda_{1,t} = 0.1 + 0.32 = 0.42$$

and for detecting a one half standard-deviation change,

$$\lambda_{1,t} = 0.1 + 0.16 = 0.26$$

An overall probability of 0.05 was desired for an alert, under the null hypothesis of no change in the visit probabilities. In addition, because 287 CUSUMs are being tested simultaneously, adjustment is needed for multiple testing (because  $287 \times 303$  values of the CUSUM are examined). A Bonferroni adjustment can be made by using  $287 \times 303$  instead of 303 in the run-length calculations. In particular, because run lengths have an exponential distribution (17),  $p(\text{run length} < 287 \times 303) = 1 - \exp(-287 \times 303 \times \mu) = 0.05$ , which implies an average run length of  $1/\mu = 1,695,366$ .

Next, the value of the tract-specific threshold ( $h$ ) that is consistent with this average run length and with the tract-specific values of  $\lambda_0$  and  $k$  was determined by using an algorithm described elsewhere (13). Then, time-varying tract-specific values of  $h_t$  were determined by either of the following methods:

1. If  $\lambda_{0,t}$  was close to any of the average values of  $\lambda_0$ , the associated value of  $h$  was adopted; if not,
2. A regression equation relating  $h$  and  $\lambda$  was estimated by using the 287 average values of  $\lambda_0$  and the 287 associated values of  $h$ . The regression equation was

$$h = 8.18 + 32.04\lambda$$

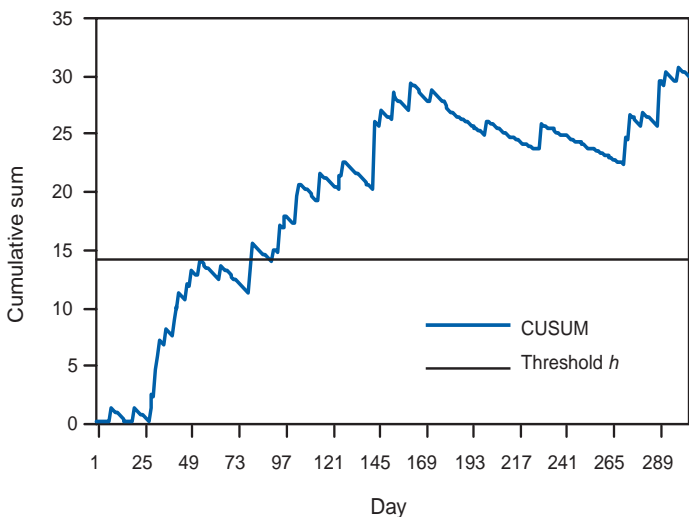
The Poisson CUSUM (equation 2) was then started for each tract on January 1, 1999, by using the observed number of daily visits, the expected number of daily visits ( $\lambda_{0,t}$ ), and values of  $h$ ,  $h_p$ ,  $k$ , and  $k_p$ , as described previously.

## Results

Of 287 census tracts, 58 (19%) had  $\geq 1$  signal during the 303-day monitoring period. In 19 (37%) tracts, the signals were short-term and continued no longer than 30 days. Of the remaining 39 tracts with signals, the majority were either sustained for approximately the latter half of the monitoring period (12 tracts) or characterized by a rapid increase in the CUSUM near the end of the monitoring period (14 tracts).

Tract 26 had an average 0.111 cases/day during the 3-year base period, which increased to an average of 0.145 cases/day during the monitoring period (Figure 1). The initial increase in the CUSUM began in late January. Cases were observed on January 28, 29, and 31; additional cases were observed on February 1, 2, and 3. Thirteen cases were observed during a 27-day period that began on January 28, for an average of 0.481 cases/day, substantially higher than the baseline of 0.111 cases/day. The CUSUM continued to increase until June, indicating a sustained period of higher-than-average visitation rates, and then declined slightly until September.

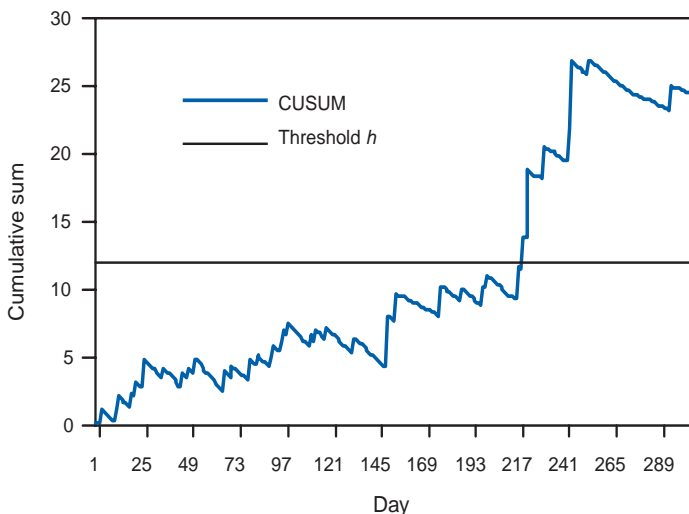
**FIGURE 1. Cumulative sum (CUSUM) chart for lower respiratory infection episodes — Census Tract 26, Boston, Massachusetts, January–October 1999**



During the base period, tract 83 had an average of 0.120 cases/day; this rose to 0.135 cases/day during 1999 (Figure 2). Cases leading to the alert occurred on August 4, 6, and 9 (two cases were observed on August 9). These four cases in 6 days (0.67 cases/day) were sufficient to generate an alert, particularly because the CUSUM had been increasing slowly during the preceding months.

During the calibration period (1996–1998), 33.4 cases/day occurred in the study region; during the first 303 days of 1999, an average of 36.8 cases/day occurred. The daily increase was >10%, and this is easily picked up by the CUSUMs in multiple subregions.

**FIGURE 2. Cumulative sum chart for lower respiratory infection episodes — Census Tract 83, Boston, Massachusetts, January–October 1999**



## Results for Monitoring Regional Neighborhoods

Neighborhoods consisting of each individual region and its immediately adjacent neighboring regions were monitored to illustrate the surveillance of local regional statistics. Of 287 census tracts, 173 (60%) had at least one signal during the monitoring period, and 43 also signaled under the original Poisson CUSUM. Among the 173 signaling tracts, 90 (52%) sustained signals for the latter half of the monitoring period, and 25 (14%) witnessed rapid increases in their CUSUMs near the end of the monitoring period. The distribution of regions that had CUSUMs above the threshold on the last day of the monitoring period (i.e., day 303), under both the original Poisson CUSUM and the local statistics CUSUM, is illustrated (Figure 3). More regions signal when the local statistic is used; here the search for spatial patterns occurs on a broader geographic scale. The northern, southwestern, and southeastern portions of the study area emerge as subareas that deviate substantially from baseline expectations established during 1996–1998.

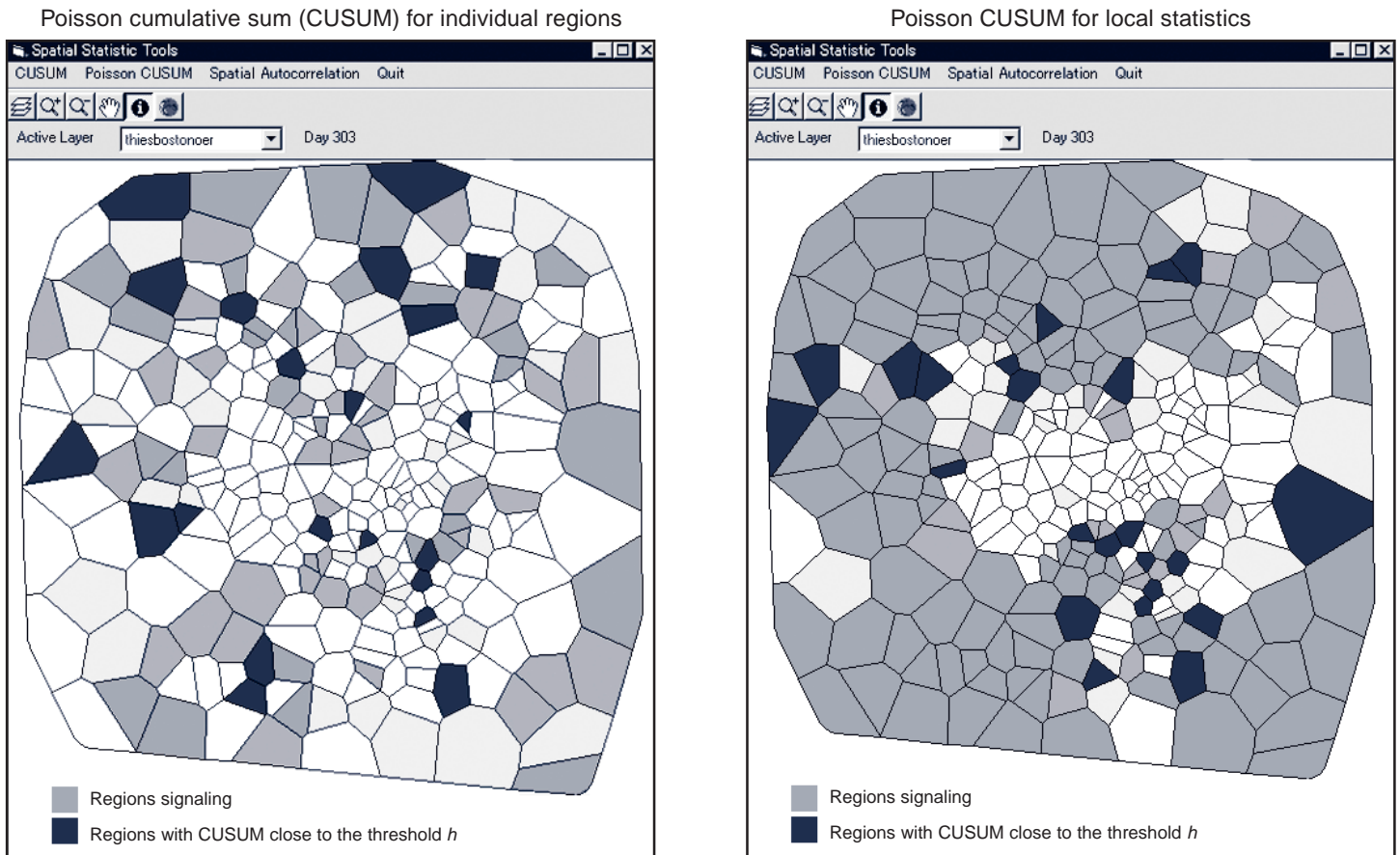
The statistical significance of the local statistics was derived by using a Bonferroni correction for the number of regions. This is conservative because the local statistics are correlated. Monte Carlo simulations were conducted by using 30- and 100-region subsets of the original study area to determine more appropriate thresholds for the local statistics CUSUM. To achieve a 0.05 probability of a false alert during the 303-day monitoring period, the target  $ARL_0$  under the null hypothesis can be calculated by using

$$p(\text{run length} < 303 \times n) = 1 - e^{(-303 \times s \times \mu)} = 0.05$$

where  $n$  is the number of regions and  $\mu = 1/ARL_0$ . For multiple values of  $s$ , the target  $ARL_0$  was calculated and the corresponding CUSUM parameters were obtained. The false-alert rates obtained by the simulations under the null hypothesis are provided (Table 2). Apparently, the appropriate value of  $s$  is 50%–60% of the number of regions  $n$  when the neighborhood is defined by the binary adjacency described previously. Using different definitions of the neighborhood would change the appropriate value of  $s$ .

On the basis of this result, local statistics CUSUM analysis was conducted on the Boston data by using  $s = 160$ , which is approximately 55% of the total number of tracts. This time, 183 census tracts, 10 tracts more than before, had at least one signal during the monitoring period, but no change was noted in terms of the day and the tract of the first signal.

**FIGURE 3. Distributions of regions that signaled on day 303 of the monitoring period, indicating lower respiratory infection episodes — Boston, Massachusetts, January 1–October 30, 1999**



**TABLE 2. False-alert rates\* simulated under the null hypothesis**

30 regions			100 regions		
$s^\dagger$	$ARL_0^{\S}$	Signaling probability	$s$	$ARL_0$	Signaling probability
30	177,216	0.024	100	590,720	0.035
20	118,144	0.038	60	354,432	0.048
15	88,608	0.055	50	295,360	0.056
10	59,072	0.075			

\* Average: >4,000 trials.

<sup>†</sup>  $s$  = number of effectively independent regions.

<sup>§</sup>  $ARL_0$  = average run length, or time between false-alerts under the null hypothesis.

## Discussion

This paper demonstrates how the Poisson CUSUM can be used in the context of spatial surveillance. In particular, it focuses on two developments: 1) an extension to allow the use of Poisson CUSUM methods when expectations vary over time, and 2) an extension along lines originally discussed previously (3) that permits monitoring of CUSUMs in subregions and their surrounding neighborhoods. Software for the

Poisson CUSUM method is available at <http://wings.buffalo.edu/~rogerson>.

An important question raised by the implementation of these methods in the context of public health surveillance is whether accurate expectations of disease counts can be formed. To the extent that expected counts are not well-modeled, the CUSUM tends to increase, and alerts caused by deviations from expectations will be attributable more to inability to model expectations and less to any real public health problem.

The methods are ultimately better suited for certain public health problems than for others. For example, for certain biologic agents, a single case is sufficient to generate an alert, and a sophisticated statistical system is not needed. In other situations, monitoring symptoms might reveal patterns that would otherwise remain hidden in the data. In the 1993 gastroenteritis outbreak in Milwaukee, a substantial number of cases went unnoticed for an extended period (18); quick detection of spatial patterns in symptoms might have allowed a quicker public health response.

### Acknowledgments

This work was funded by National Institutes of Health Grant 1R01 ES09816-01 and National Cancer Institute Grant R01 CA92693-01. Ken Kleinman (Department of Ambulatory Care and Prevention, Harvard Medical School, Harvard Pilgrim Health Care, and Harvard Vanguard Medical Group) provided the data on lower respiratory infection episodes. Two reviewers also provided helpful comments.

### References

1. Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. *J R Stat Soc A* 2003;166:5–21.
2. Farrington P, Beale AD. The detection of outbreaks of infectious disease. In: Lierl L, Cliff AD, Valleron A, Farrington P, Bull M, eds. *Geomed '97*. Stuttgart: BG Teubner, 1998.
3. Raubertas RF. An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Stat Med* 1989;8:267–71.
4. Kleinman K, Lazarus R, Platt R. A generalized linear mixed models approach for detecting incident clusters of disease: biological terrorism and other surveillance. *Am J Epidemiol* 2004;156:217–24.
5. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *J R Stat Soc A* 2001;164:61–72.
6. Rogerson P. Surveillance methods for monitoring the development of spatial patterns. *Stat Med* 1997;16:2081–93.
7. Rogerson P. Monitoring point patterns for the development of space-time clusters. *J R Stat Soc A* 2001;164:87–96.
8. Montgomery D. *Introduction to statistical quality control*. New York, NY: John Wiley, 1996.
9. Siegmund D. *Sequential analysis: tests and confidence intervals*. New York, NY: Springer-Verlag, 1985.
10. Rossi G, Lampugnani L, Marchi M. An approximate CUSUM procedure for surveillance of health events. *Stat Med* 1999;18:2111–22.
11. Lucas JM. Counted data CUSUMs. *Technometrics* 1985;27:129–44.
12. White CH, Keats JB. ARLs and higher order run length moments for Poisson CUSUM. *Journal of Quality Technology* 1996;28:363–9.
13. Hill GB, Spicer CC, Weatherall JAC. The computer surveillance of congenital malformations. *BMJ* 1968;24:215–8.
14. Weatherall JAC, Haskey JC. Surveillance of malformations. *BMJ* 1976;32:39–44.
15. Hutwagner LC, Maloney EK, Bean, NH, Slutsker L, Martin SM. Using laboratory-based surveillance data for prevention: an algorithm for detecting *Salmonella* outbreaks. *Emerg Infect Dis* 1997;3:395–400.
16. Hawkins DM, Olwell DH. *Cumulative sum charts and charting for quality improvement*. New York, NY: Springer, 1998.
17. Brook D, Evans DA. An approach to the probability distribution of CUSUM run length. *Biometrika* 1972; 59:539–49.
18. Eisenberg JN, Seto EY, Colford JM Jr, Olivieri A, Spear RC. An analysis of the Milwaukee cryptosporidiosis outbreak based on a dynamic model of the infection process. *Epidemiol* 1998;3:228–31.