# Evaluation of Alerting Sparse-Data Streams of Population Healthcare-Seeking Data

**Howard Burkom, Yevgeniy Elbert**
*The Johns Hopkins University Applied Physics Laboratory*

**Objective**
This presentation discusses the problem of detecting small-scale events in biosurveillance data that are relatively sparse in the sense that the median count of monitored time series values is zero. Research goals are to understand conditions when methods adapted for sparseness are warranted, to examine adaptations of control charts and other algorithms under these scenarios, and to compare the detection performance of these algorithms.

**Background**
Biosurveillance systems have had little success detecting community-level outbreak events using broad syndrome definitions. The growing use of more specific information such as electronic medical records allows a focus on smaller sets of more relevant cases and emphasizes the need to monitor sparse time series. Much has been published on the biosurveillance data modeling to account for systematic or cyclic behavior, but little on the monitoring of sparse series. Literature may be found in statistical process control applied to chronic disease surveillance [1], but the infectious disease data environment poses additional challenges of monitoring transient events, instead of mean shifts in incidence levels, and changing background behavior. Some monitoring institutions aggregate data streams to improve predictive model performance and to avoid multiple testing problems, but at the cost of missing weak signals. Others apply simple data thresholds to sparse data. The current effort investigates benefits realizable by statistical monitoring of sparse series.

**Methods**
We measured the effectiveness of sparse time series methods on both authentic and simulated data streams. The authentic data streams were small-scale series of 1400 days of daily visit counts in 20 syndromic categories. The simulated data streams were sets of 15000 draws from Poisson distributions with the lag-1 autocorrelation coefficient $\tau$ controlled to be zero, to simulate independent data, or a multiple of 0.1, to examine the effect of temporal dependence. Alerting algorithms considered include adaptations to CUSUM and Xbar charts, temporal scan statistics adaptations, and the simple data threshold. Evaluation software was built to facilitate investigation of parameter changes and inclusion of other methods. These evaluations measure the sensitivity, at manageable background alert rates, to plausible, stochastic (lognormal or uniform) signals representing the data effects of simulated outbreaks.

**Results**
As a sample result, Table 1 gives the sensitivity of 5 methods at an alert rate of 1 per 12 weeks for 1000 trials of a stochastic uniform signal of 10 attributable cases over a spread of 14 days. An adaptive CUSUM-based chart gave consistent performance across data types, while for data streams with mean below 0.5 counts/day and $\tau < 0.1$, the temporal scan statistic gave a clear sensitivity advantage. Poisson regression analysis was used to summarize the effects of the data mean and $\tau$, signal type, signal size, false alarm rate, and alerting method on sensitivity. Highly autocorrelated data streams presented a problem for separating weak signals from natural association.

**Conclusions**
For sparse data streams that were not highly autocorrelated, sensitivity improvements of 10-20% were typical for the adapted alerting methods over the simple threshold method sometimes employed. This improvement was consistent for the authentic data streams and is magnified with the number of regions and subsyndromes monitored. Such improvements must be combined with other informatics advances before biosurveillance systems will detect outbreaks of modest size.

**References**
[1] Woodall W., "The Use of Control Charts in Health-Care and Public-Health Surveillance," Journal of Quality Technology 38, 2:89-104.

| method | data mean 0.2 | data mean 0.6 |
|---|---|---|
| adaptive CUSUM | 0.953 | 0.716 |
| adaptive Xbar | 0.894 | 0.484 |
| temporal scan_3 | 0.975 | 0.754 |
| temporal scan_7 | 0.941 | 0.695 |
| data threshold | 0.896 | 0.547 |

Table 1: Sensitivities at an alarm rate of 1 per 12 weeks for a 1000 trials of a stochastic uniform signal of 10 attributable cases over a spread of 14 days.