



# The Role of Data Aggregation with Applications in ESSENCE II

Howard Burkom

National Security Technology Department,  
Johns Hopkins University Applied Physics Laboratory

Yevgeniy Elbert

Walter Reed Army Institute for Research  
Global Emerging Infections System

National Syndromic Surveillance Conference  
New York Academy of Medicine  
New York, NY October 24, 2003



# Problem Statement



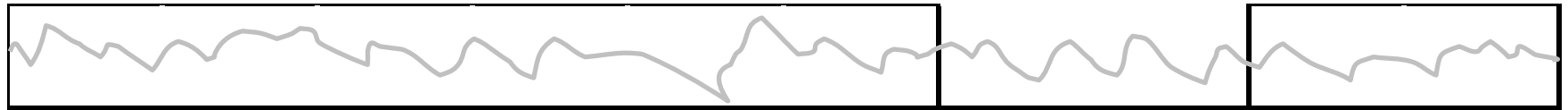
- Working definition of syndromic surveillance: the monitoring of available data sources for outbreaks of unspecified disease or of specified disease before the confirmation of identifying symptoms
- Objective: to complement existing sentinel surveillance by identifying outbreaks with false alarm rates acceptable to the public health infrastructure.
- Key question: How should we filter, aggregate data for monitoring?
  - What covariates should we include?
  - How should we monitor multiple data streams?



# Aggregating Data in Time



Data stream(s) to monitor in time:



## baseline interval

Used to get some estimate of normal data behavior

- Mean, variance
- Regression coefficients
- Expected covariate distrib.
  - spatial
  - age category
  - % of claims/syndrome

## guardband

Avoids contamination of baseline with outbreak signal

## test interval

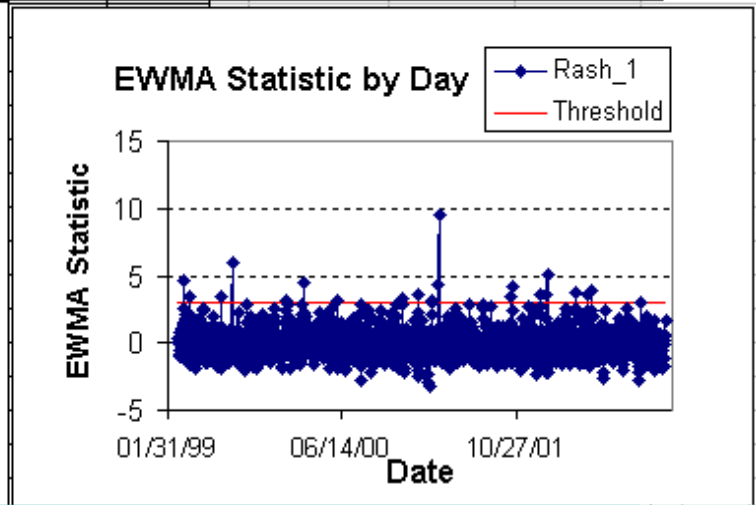
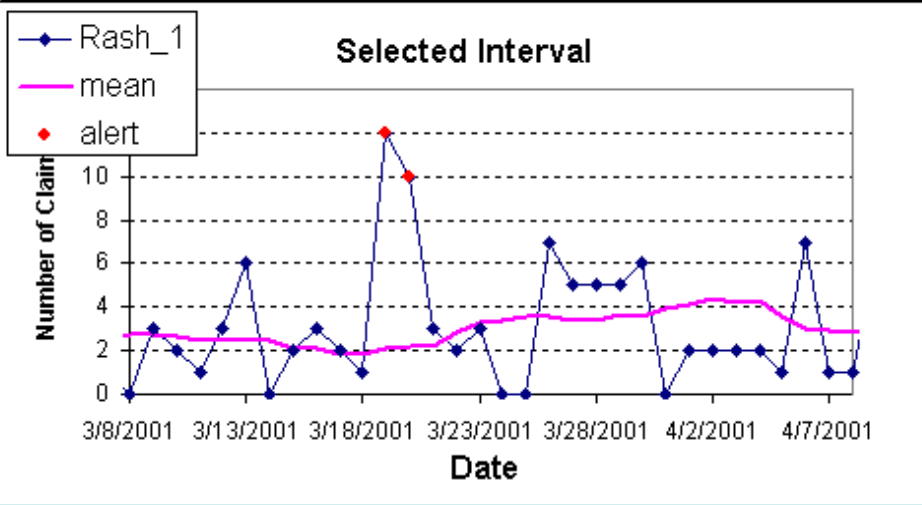
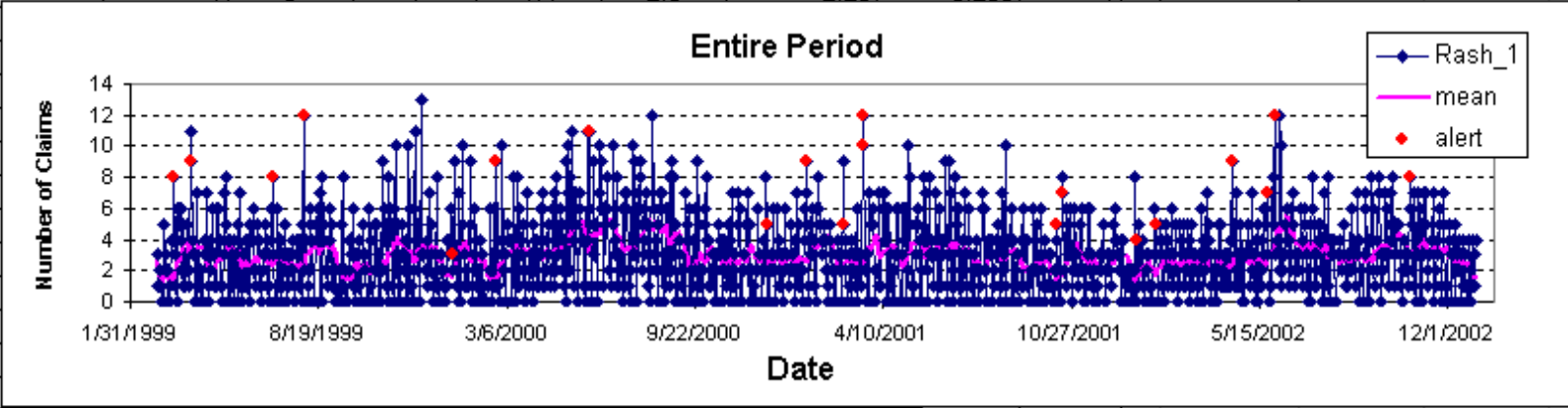
- Counts to be tested for anomaly
- Nominally 1 day
- Longer to reduce noise, test for epicurve shape
- Will shorten as data acquisition improves



# Syndromic Data: Algorithm Performance



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Rash_1	Outbreak Severity	Baseline Lag	Baseline Length			Reset Flag	Smoothing Coefficient	Threshold			Selected Interval			
26	0.0	2	14			1	0.40	3			3/8/2001	36958		
app_date	Outbreak Shape	Rash_1	Added cases	Data	Sigma Estimate	Mean Estimate	Smoothed	Statistic	alert?		4/8/2001	36989		
3/1/1999		3	0	3	1.0	3.0	3.00	0.000	-1	3				
3/2/1999		1	0	1	1.4	2.0	2.20	0.283	-1					

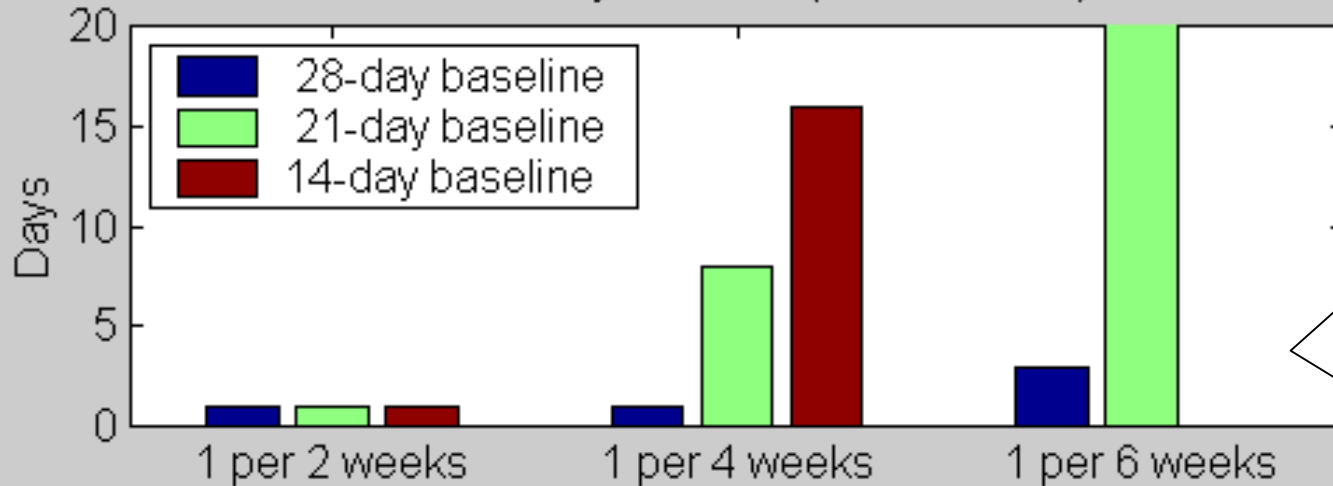




# Algorithm Performance Assessment

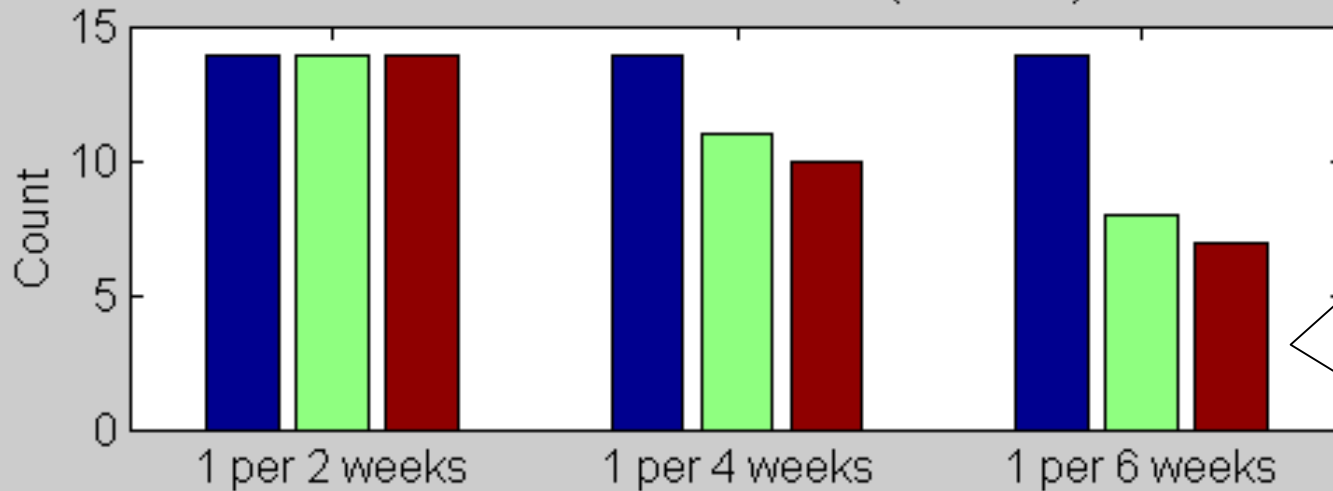


Median Days To Alert (15 outbreaks)



Timeliness relative to specificity (relevant snapshots of AMOC curve)

Number of Events Alerted (out of 15)



Sensitivity relative to specificity (relevant snapshots of ROC curve)

False Alert Rate



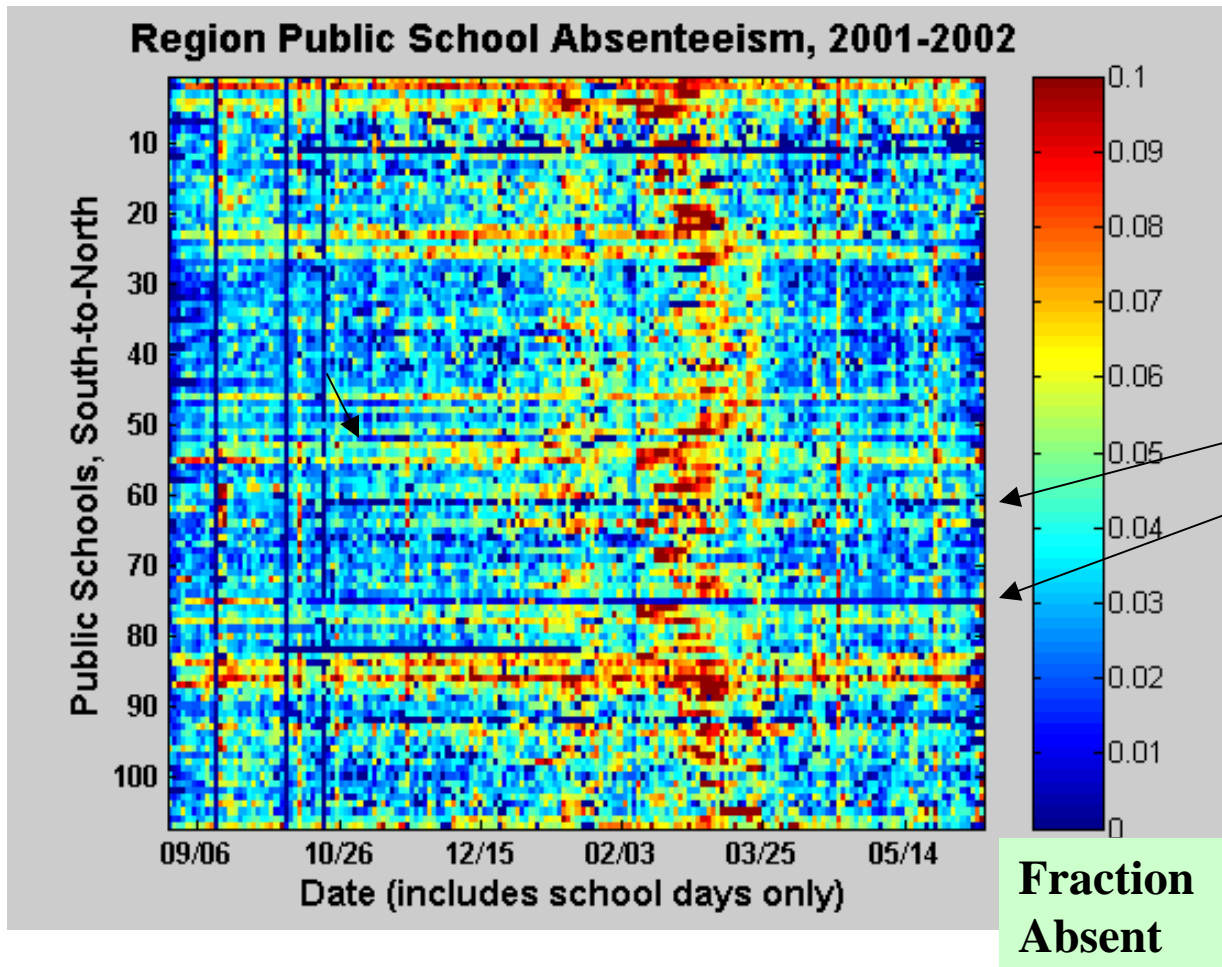
# Spatial Aggregation: Scan Statistics



- Use of spatial (& spatiotemporal) scan statistics to monitor for outbreaks
  - Find approx. location, extent of space-time interaction
  - Accounts for expected spatial inhomogeneities
- Key step in controlling for inhomogeneity—determining the expected spatial distribution
  - Surveillance distribution often not population-based
  - Dependence on provider, consumer concentrations
  - Incidence may be nonuniform
- In practice, sliding temporal baseline approach often used to infer distributions



# Data Analysis: Quality & Consistency



Questionable,  
inconsistent  
reporting can  
raise  
false alarm level



# Statistical Process Control Approaches

- Applied to data or residuals
- Univariate
  - Exponential weighted moving average (EWMA)
  - CUSUM-based (CDC EARS)
  - For multiple univariate: combined inference using simple OR, Edgington comb., Bayes Belief Nets
- Multivariate
  - Hotelling's  $T^2$  (used in cyberattack detection)
  - Crosier's shrinking CUSUM
  - Lowry's MEWMA
  - Modifications to reduce alerts due to irrelevant changes in data relationships



# MSPC References



1. Ye, N., Cheng, Q., Emran, S, Vilbert, S, Hotelling's  $T^2$  Multivariate Profiling for Anomaly Detection, *Proceedings of the 2000 IEEE Workshop on Information Assurance and Security*, West Point, NY, June 2002.
2. Lowry, C.A., Woodall, W.H., A Multivariate Exponentially Weighted Moving Average Control Chart, *Technometrics*, February 1992, Vol. 34, No. 1, 46-53
3. Crosier, R.B., Multivariate Generalizations of Cumulative Sum Quality-Control Schemes, *Technometrics*, August 1988, Vol.. 30, No. 3



# Covariance-Based MSPC



- $X$  = multivariate data from test interval
  - Multiple time series: syndromic claims, OTC anti-flu sales, etc
  - Single day, or weighted mean (Ye used EWMA)
- $\mu$  = vector mean estimated from baseline interval
- $S$  = estimate of covariance matrix calculated from baseline interval
- $T^2$  statistic:  $(X - \mu) S^{-1} (X - \mu)$
- Alert based on empirical distribution of  $T^2$  for controlled false alarm rate



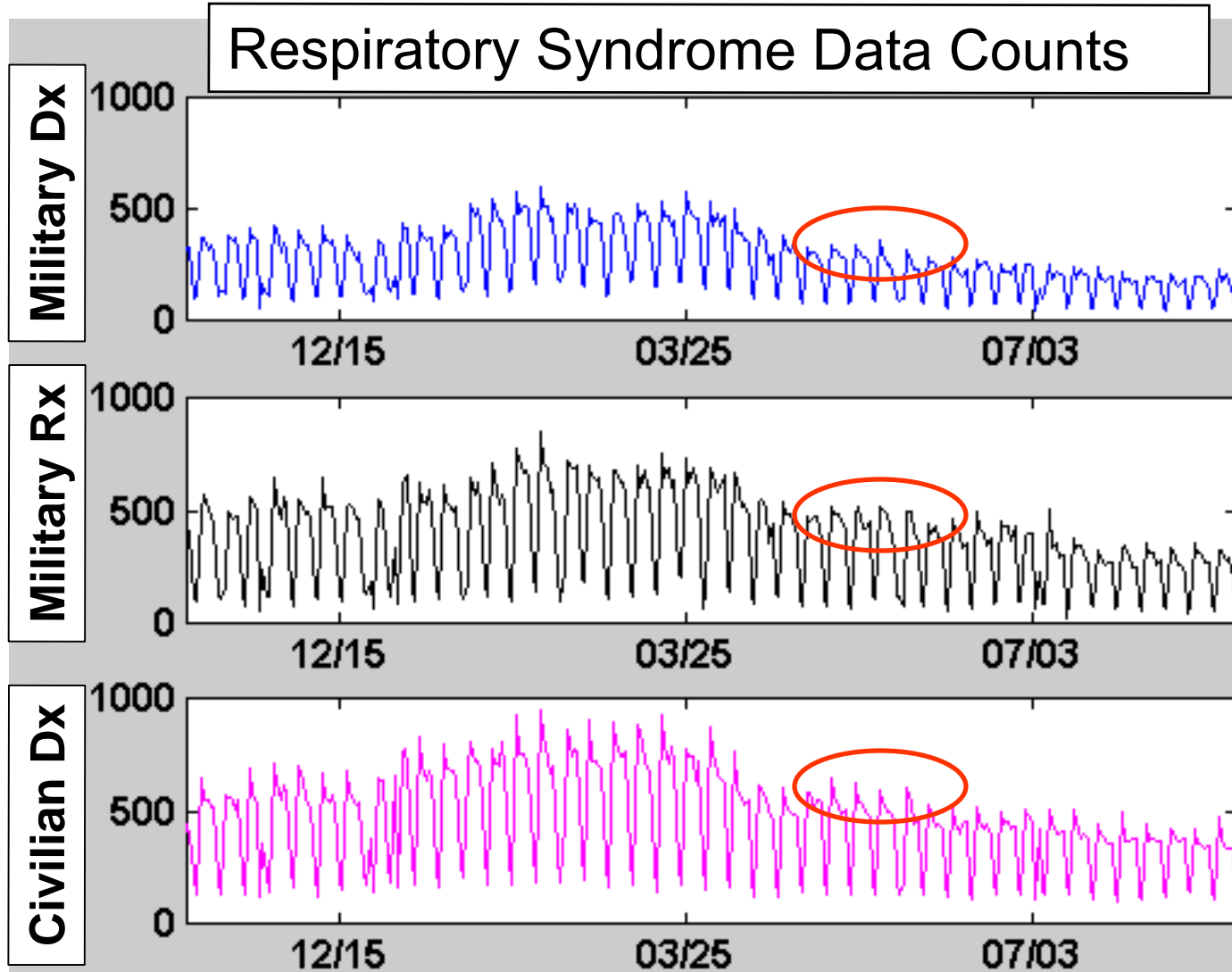
# Multivariate Data

## Surveillance: Recent DARPA Exercise

- Search for authentic outbreaks
- 2 years of daily data from 5 cities
- 3 data sources from each city
  - Military outpatient clinic visits
  - Civilian physician office visits
  - Military prescriptions
- 15 outbreaks chosen by committee of epidemiologists & physicians
  - Syndrome groups: Respiratory, gastrointestinal
  - Start date, PH recognition date, peak date, end date
- Objective: timely alerts at specified false positive rates



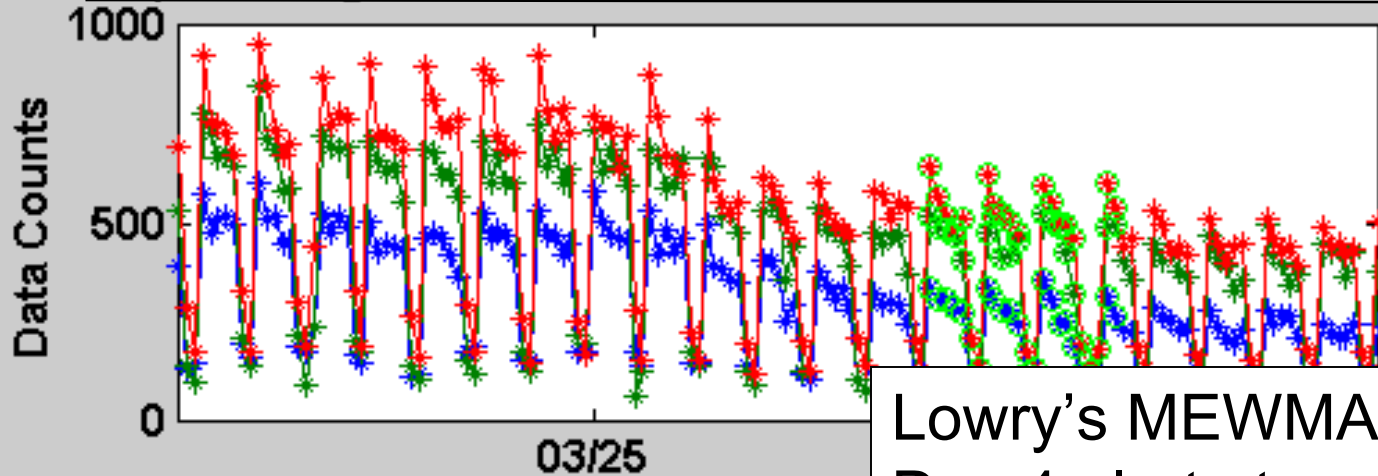
# Detection Challenge: faint rise in all 3 data sets



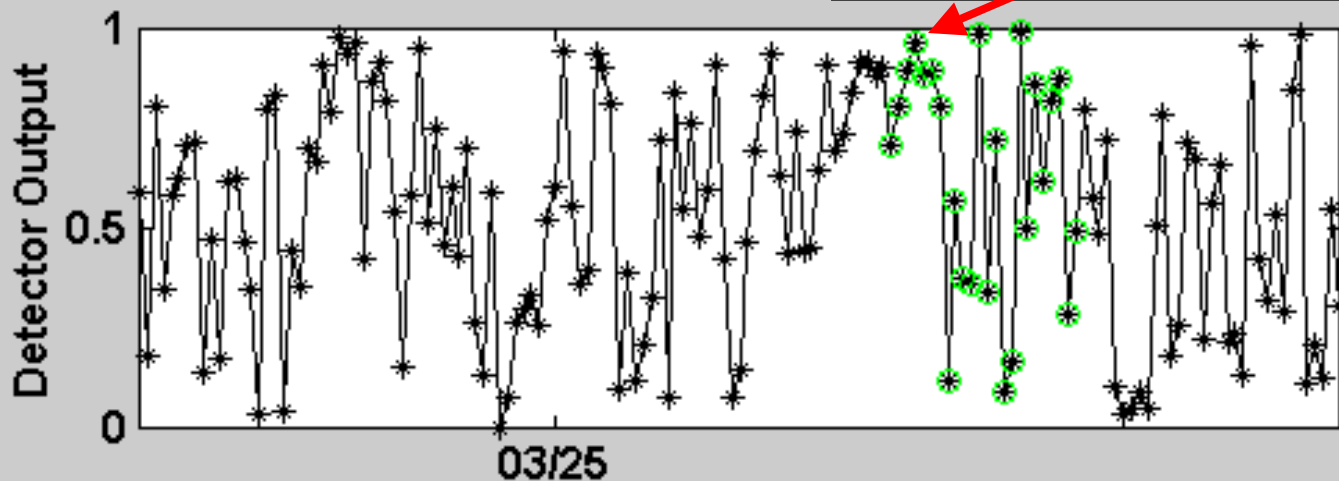


# Detection Challenge: faint rise in all 3 data sets

Respiratory Syndrome Data Counts



Lowry's MEWMA:  
Day 4 alert at each FA rate





# Managing Systematic, Unexpected Effects: Provider Count Regression

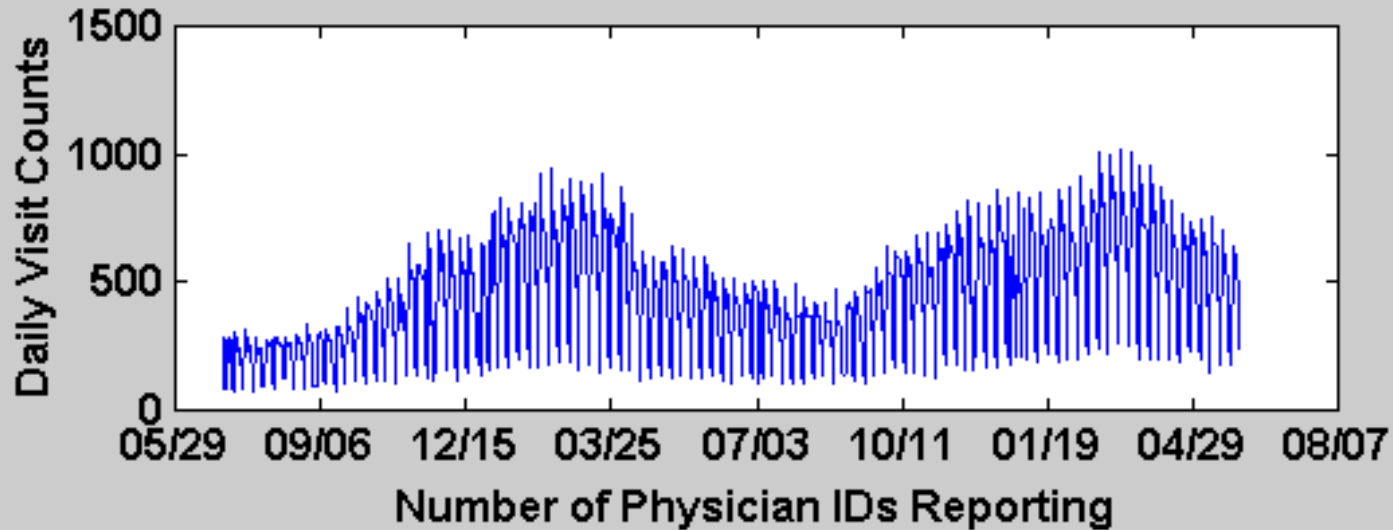
- Concept:
  - tabulate # doctors, clinics, pharmacies
  - use residuals of lin. regression of daily data counts on # providers
  - accounts for known & unknown dropoffs by computing actual counts vs expected, given daily # providers
- Can apply process control algorithms to residuals
- Significantly attenuates day-of-week effect
- Additional utility: management of late reporting effects (better than survival analysis?)



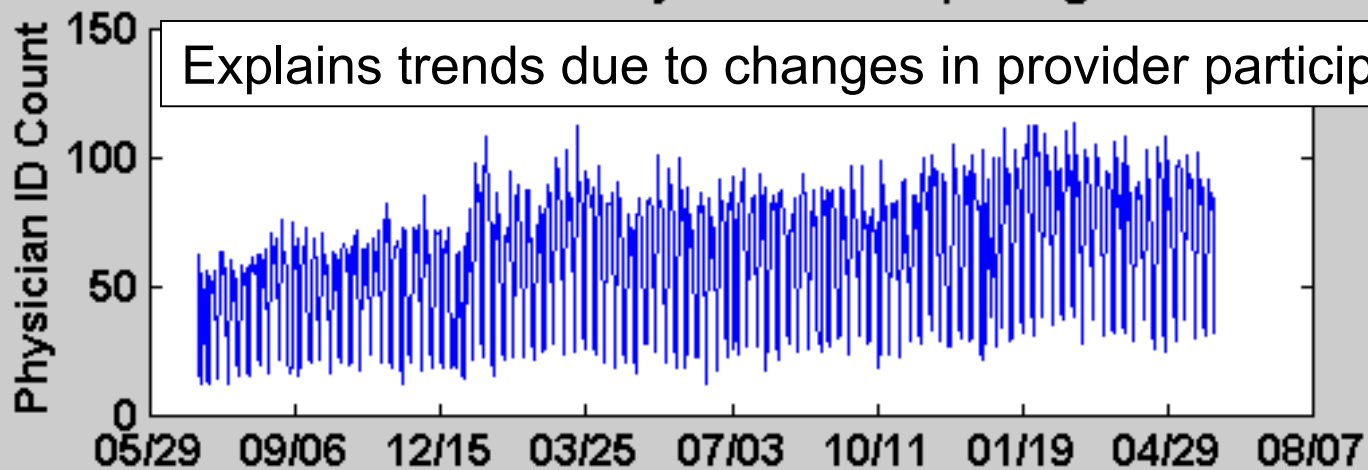
# Recent Method: Using Counts of Physician IDs



Counts of Civilian Office Visits: Respiratory Syndrome Group



Explains trends due to changes in provider participation

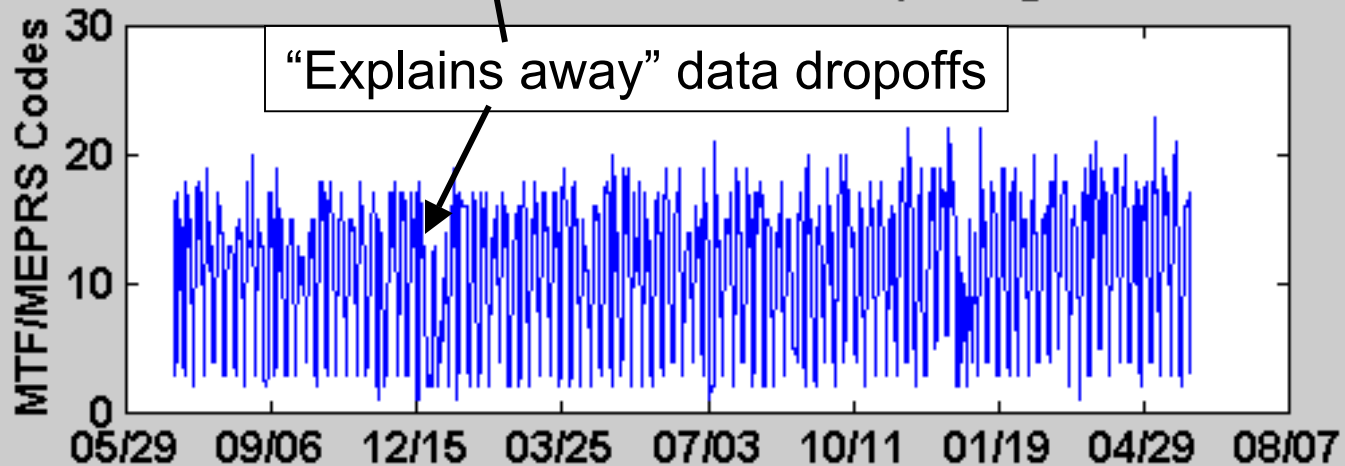
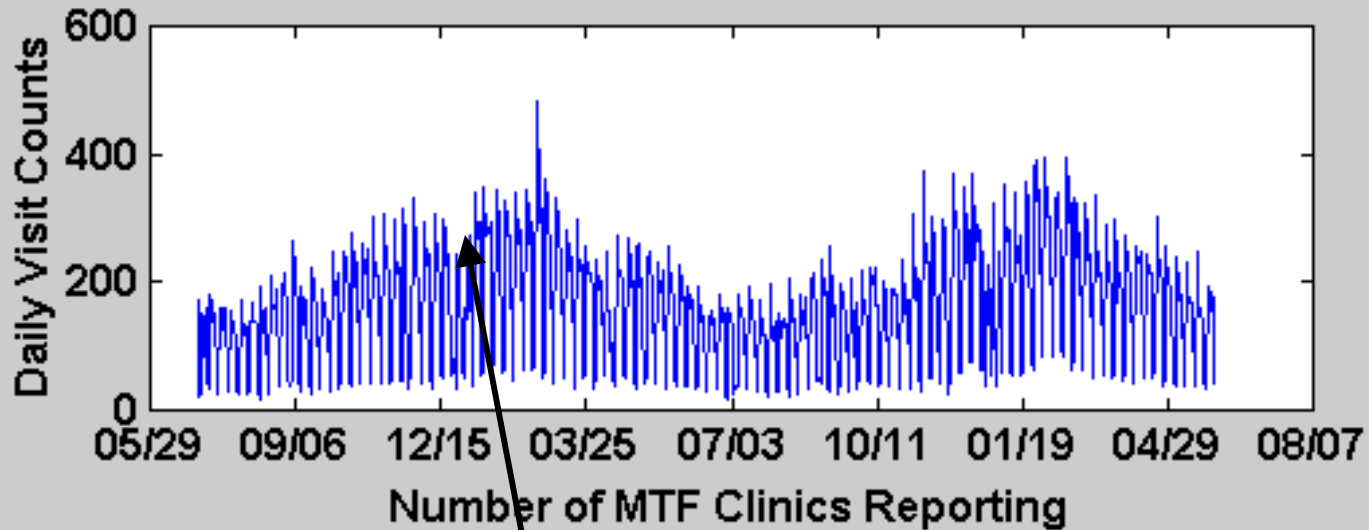




# Using Counts of Reporting Clinics



Counts of Military Clinic Visits: Respiratory Syndrome Group



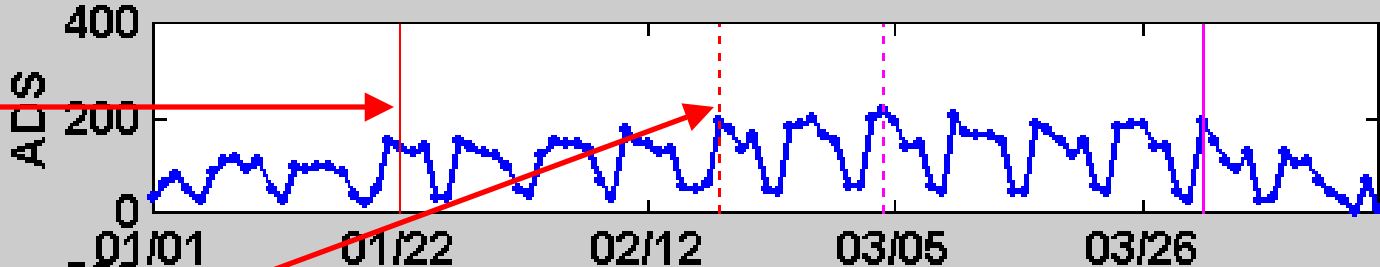


# Effectiveness of Multivariate Detection

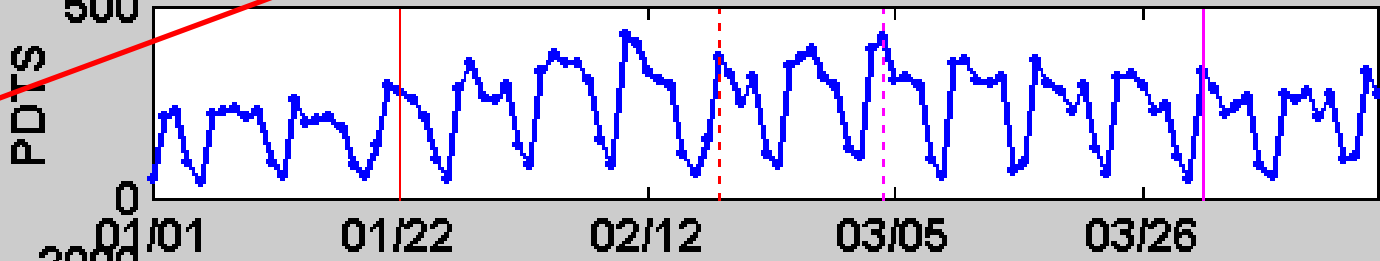


Outbreak #5 synd.: resp city: Louisville Start: 1/22/2003 PH: 2/18/2003

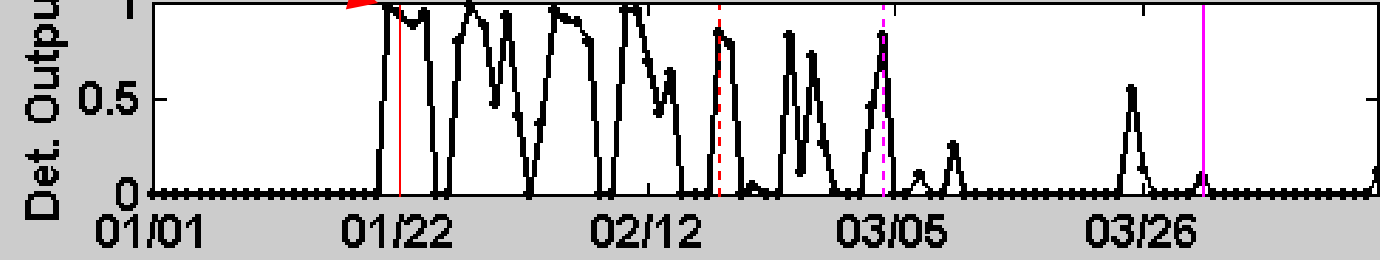
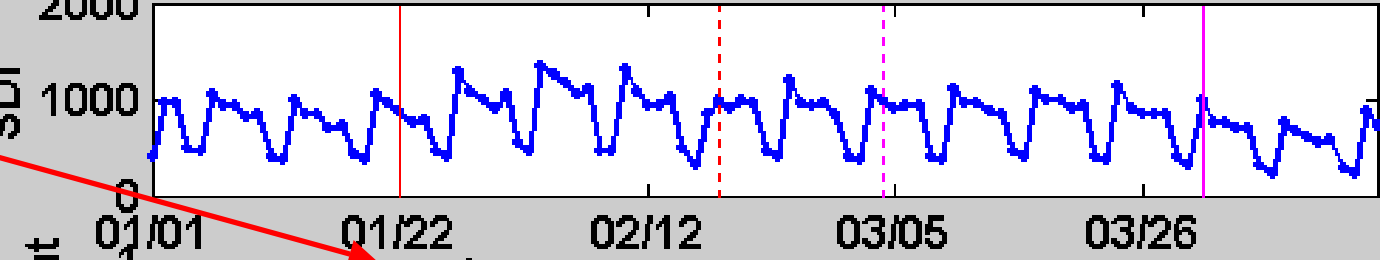
Nominal outbreak start date



Nominal Public Health awareness date

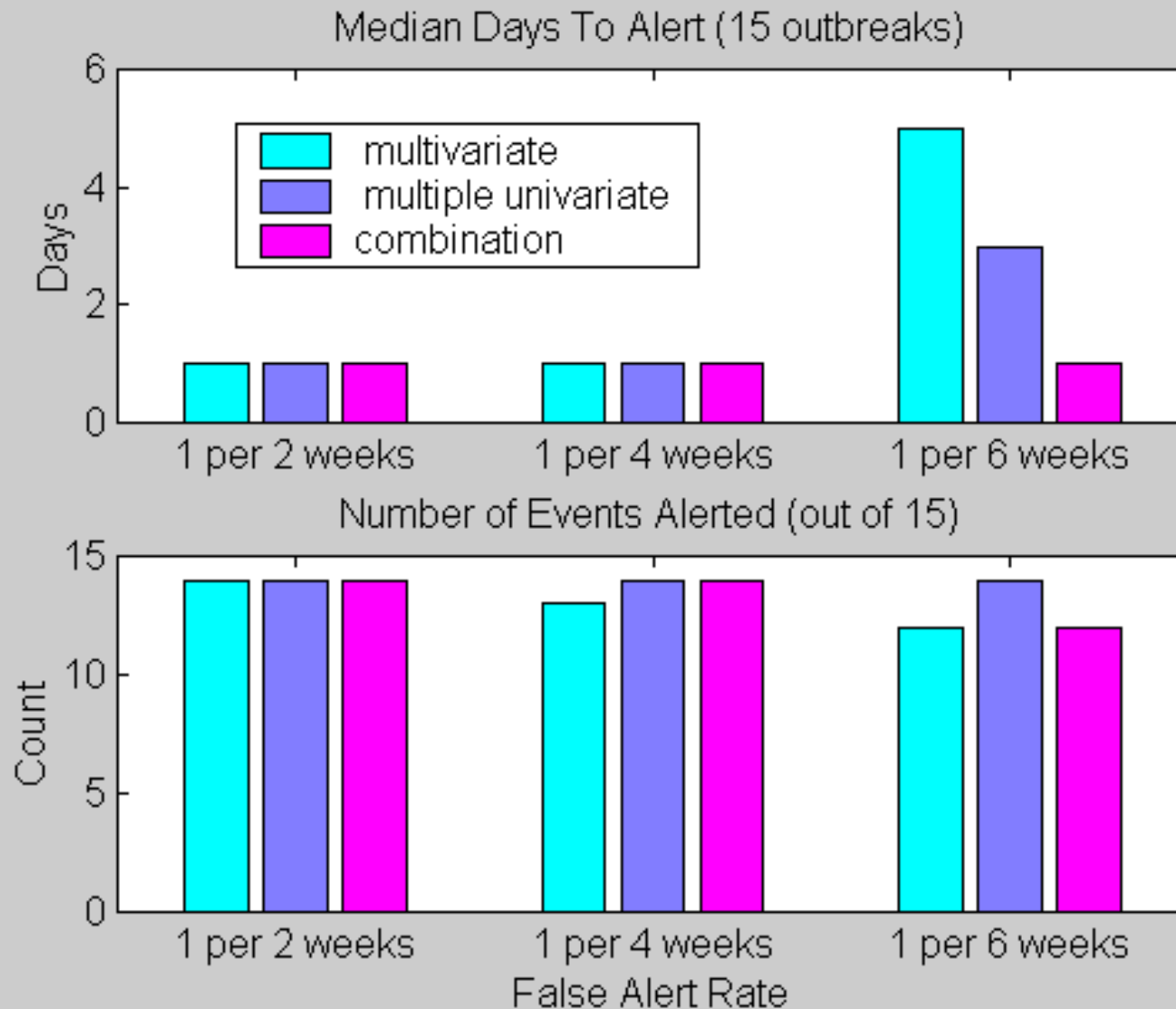


Alert day before nominal start date





# Evaluation Summary: GI Outbreaks





# BACKUPS



# Finding Clusters with Multiple Data Sources



- **For candidate cluster J1, the likelihood ratio is:**

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

where O1 = number of cases inside J1,

E1 = number of cases outside J1,

N = total case count

- **Extension by treating multiple sources as covariates:**

$$O1 = \sum O1_k, \quad E1 = \sum E1_k, \quad N = \sum N_k, \quad \text{for sources } k=1, \dots, K$$

– problem of adding sources with mismatched scales, variances

- **Alternate multisource approach: “stratified” scan statistic**

$$\sum \log( LR(J1_k) ), \quad k=1, \dots, K$$

– reduces chances for a noisy source to overwhelm others

– can cost power to detect faint signal spread over sources



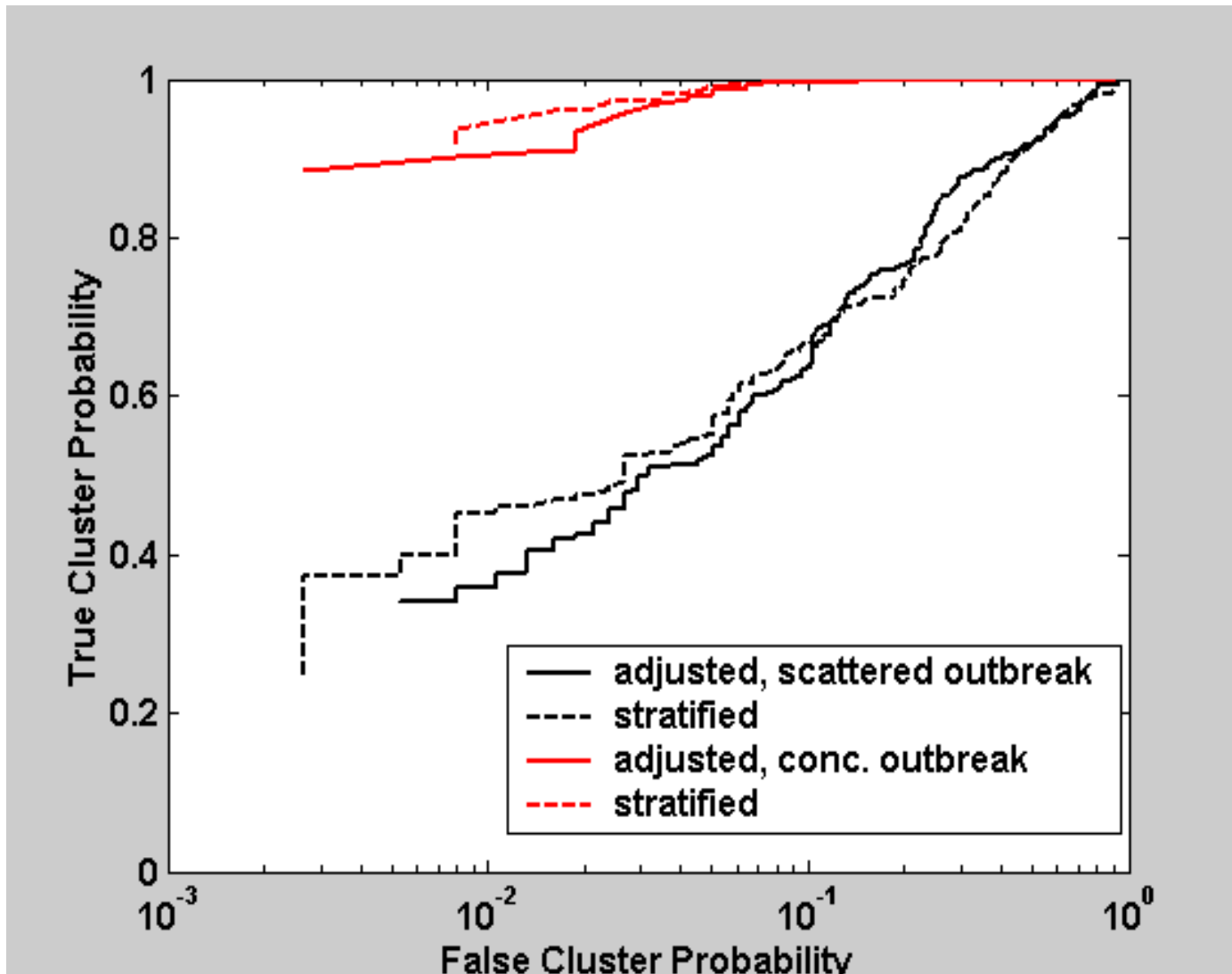
# Performance Analysis for Scan Statistics



- ROC-like measure:
  - Prob(true cluster) vs Prob(false cluster)
  - Monte Carlo trials
- Stochastic noise field generation (endemic cases)
  - Each trial: N draws from expected spatial distribution
- Stochastic signal field generation (localized epidemic to detect)
  - Assume exponential falloff with distance, circular/elliptic
- Perform scan statistic & cluster analysis for each trial



# Multisource Scan Statistics: Adjusted vs Stratified

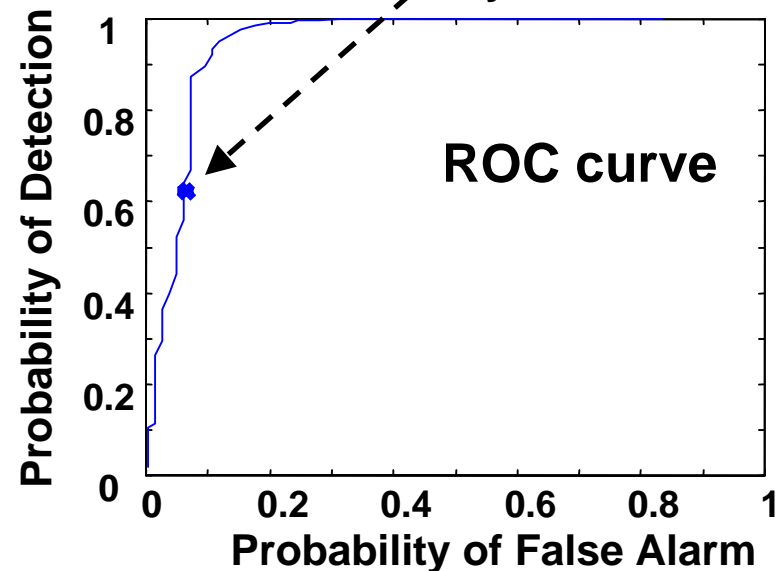
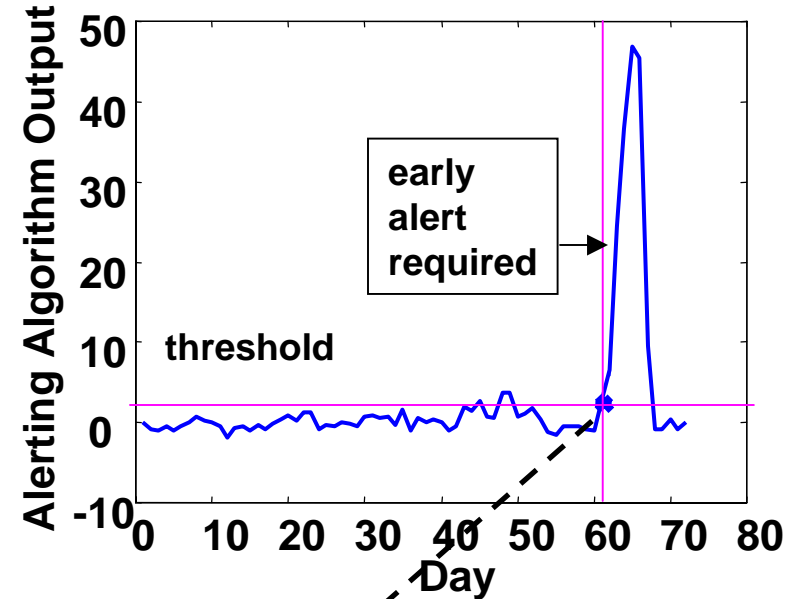




# Performance Analysis Methodology



- Seeking measures relevant to biosurveillance
- Emulating Receiver Operating Characteristic (ROC) Methods
- False alarm levels important to planning of public health follow-up
- Noise background data: authentic data or simulated based on authentic statistical moments
- Question: how to get signal representation of outbreaks?
  - Genuine outbreaks of endemic disease difficult to pinpoint
  - Using simulated epicurves based on Sartwell lognormal model



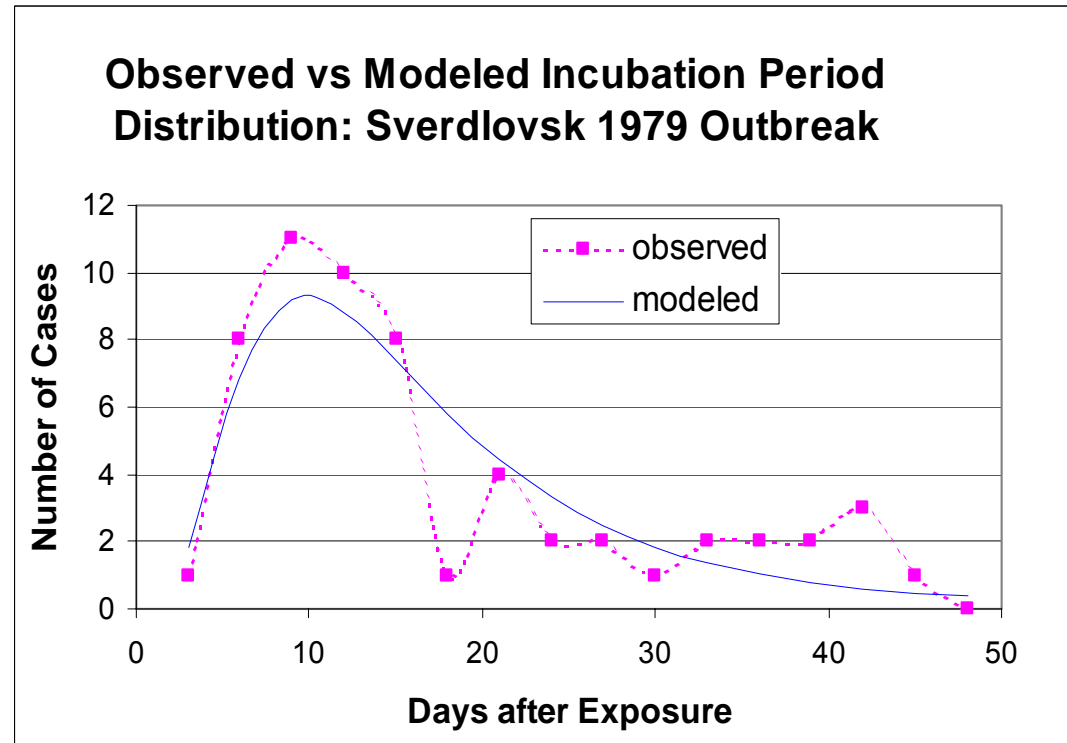


# Modeling the Signal as Epicurve of Primary Cases

- Sartwell, 1949: incubation periods lognormally distributed
- Concept to model signal as 2-parameter lognormal distribution, with density:

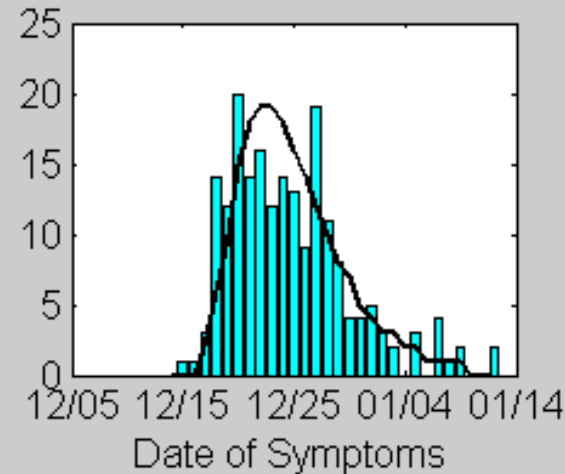
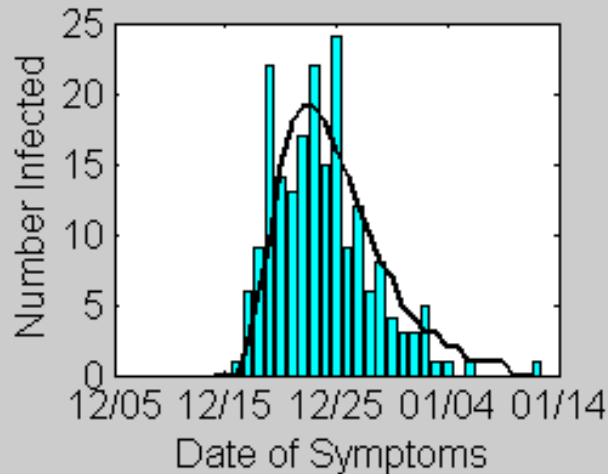
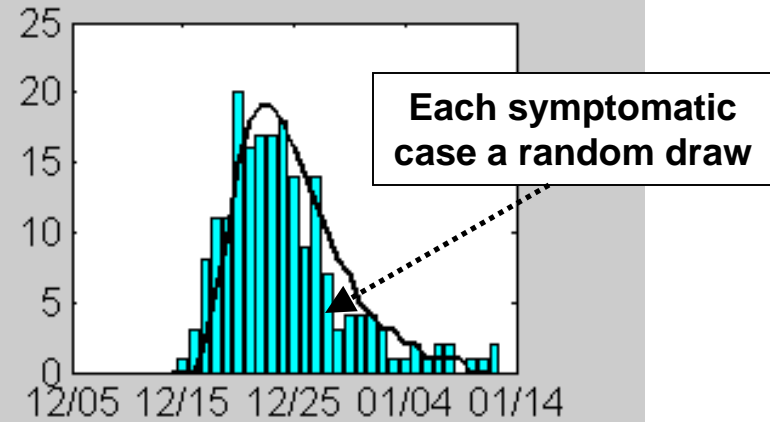
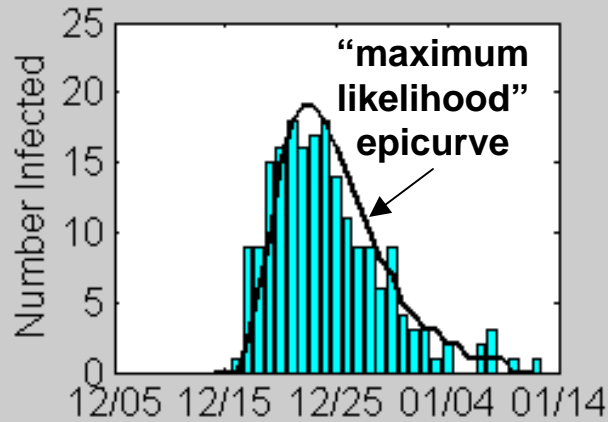
$$\frac{1}{\sigma\sqrt{2\pi x}} \exp\left(-\frac{(\log(x)-\zeta)^2}{2\sigma^2}\right)$$

- Parameters  $\sigma$  and  $\zeta$  dependent on specifics of disease
- For signal strength, set modal value =  $k$  times  $\sigma$  of data, where  $k = 1, 2, 3$  depending on application





# Signal Modeling: Realizations of Smallpox Epicurve

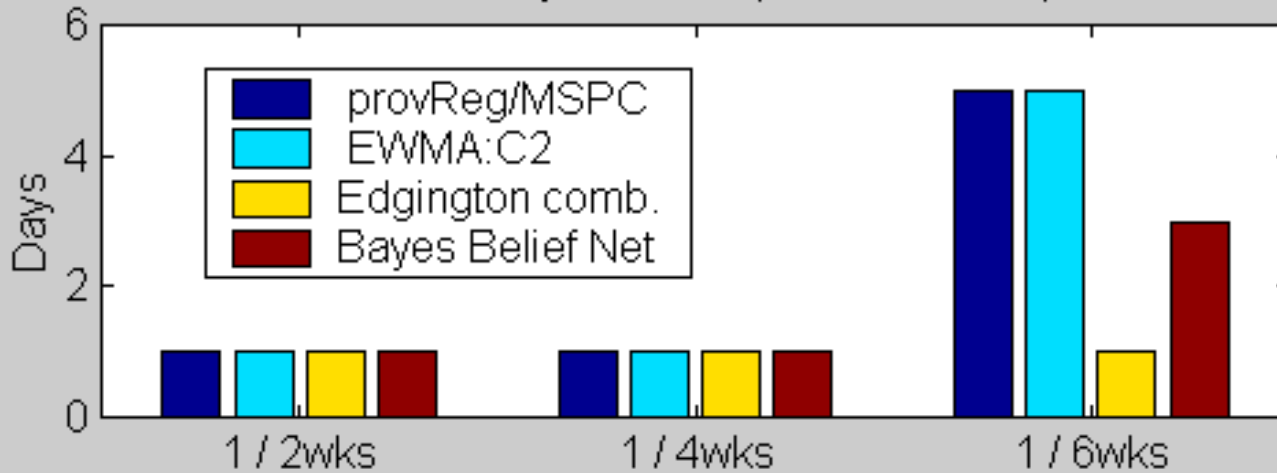




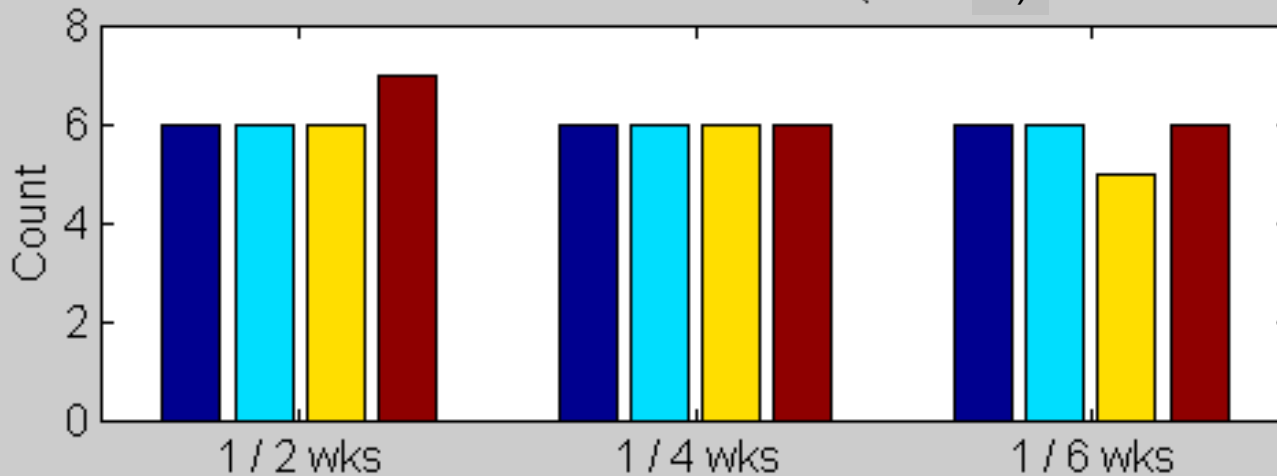
# Evaluation Summary: GI Outbreaks



Median Days To Alert (7 GI outbreaks)



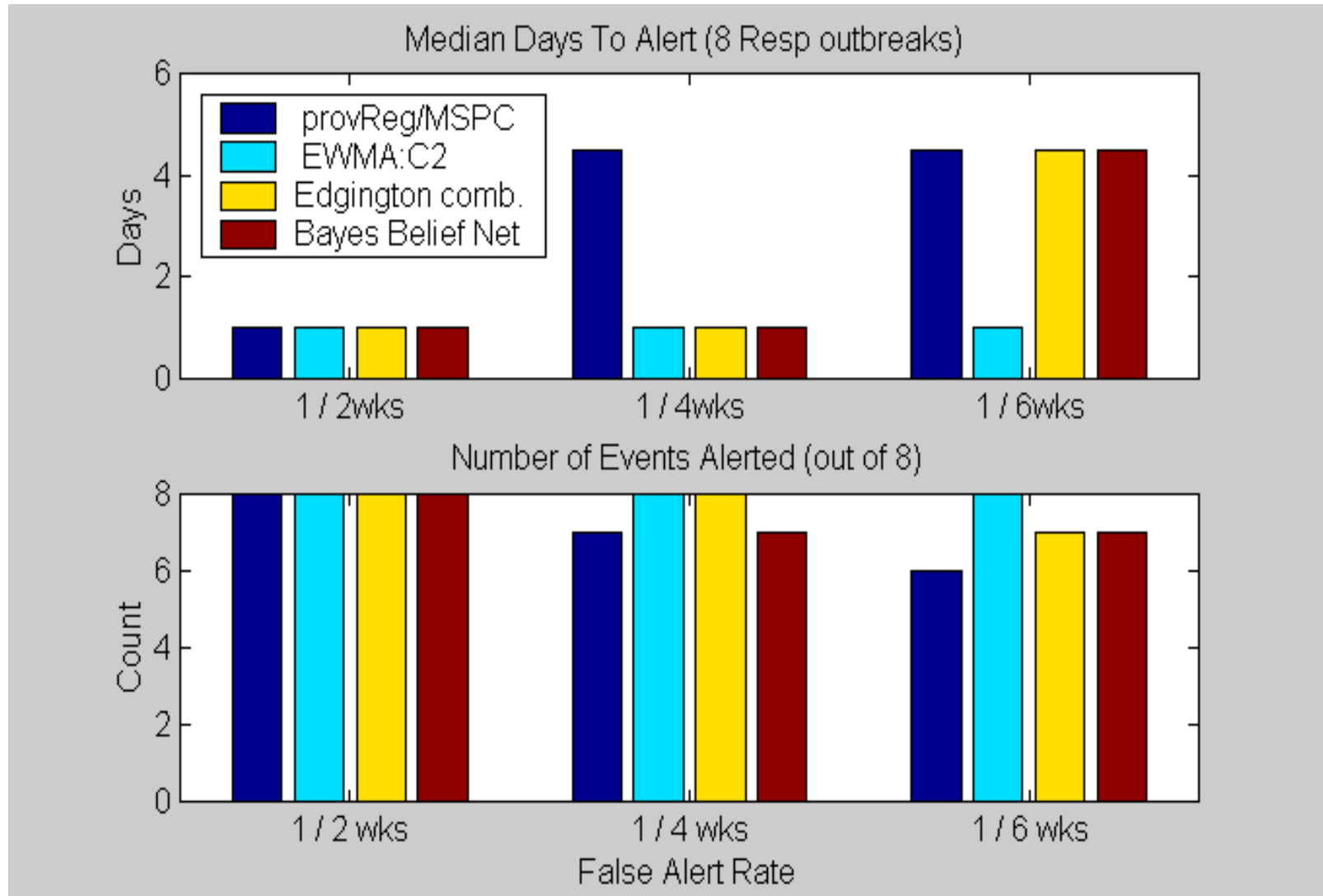
Number of Events Alerted (out of 7)



False Alert Rate



# Evaluation Summary: Respiratory Outbreaks





# Algorithm Methodologies



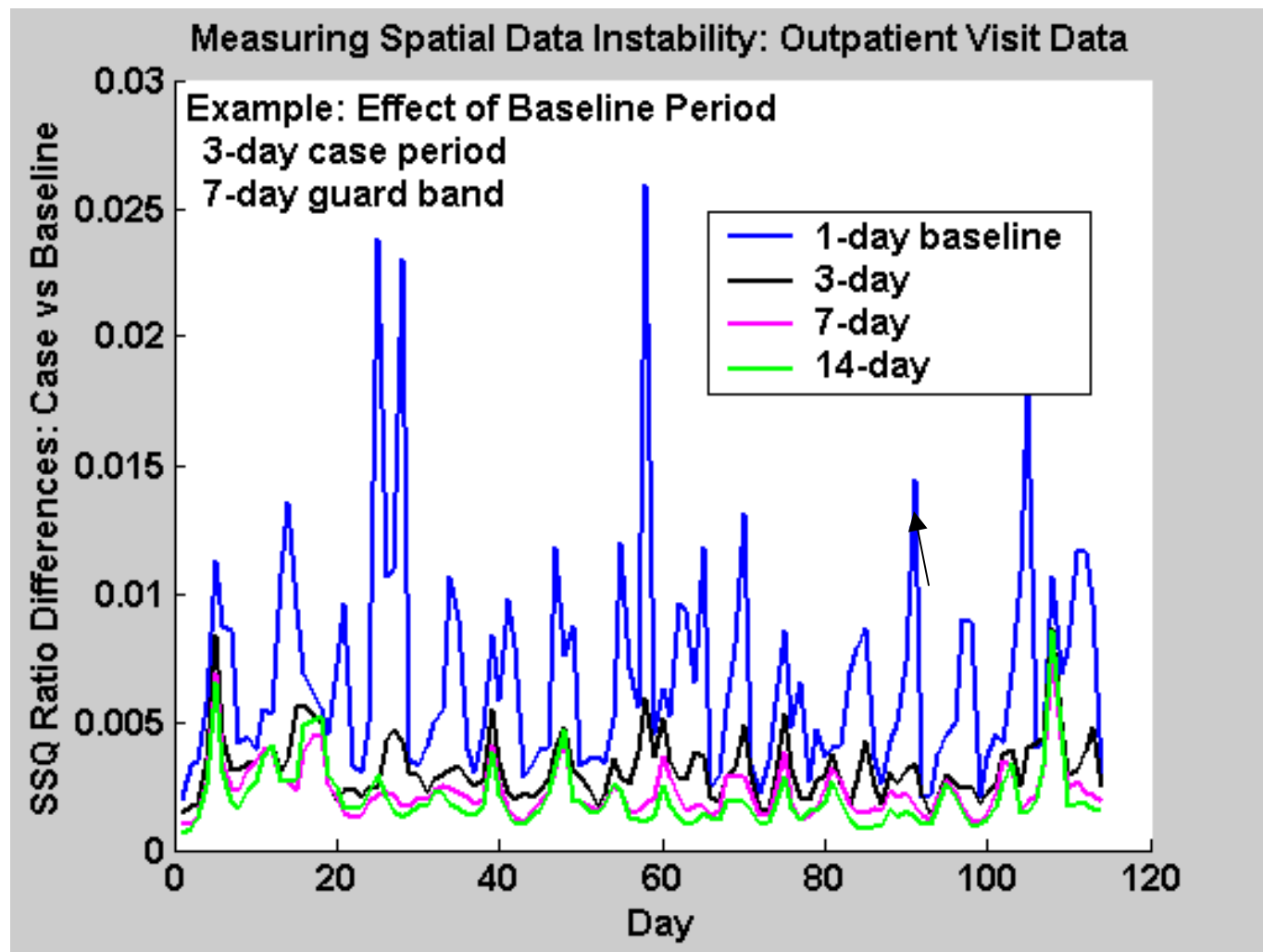
- Multivariate approach
  - Precondition data: remove weekend, holiday, other effects with regression including daily counts of providers
  - Apply MSPC: modified Hotelling's  $T^2$
- Multiple univariate approach
  - SPC on each data stream: a form of EWMA similar in effect to CDC EARS C2
  - Individual algorithm results combined using
    - Simple OR
    - Bayes Belief Net



# Analyzing Expected Spatial Distribution



- Use of baseline data to estimate spatial distrib.:
- Too short a baseline: unstable
- Too long a baseline: may not reflect recent data



$\Sigma(\text{difference of case/baseline ratios})^2$  to establish baseline period to resolve stationarity/stability issue for a representative & recent baseline