ISDS
INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Analytic Methodologies for Disease Surveillance Using Multiple Sources of Evidence

November 6, 2014

Linus Schiöler, Statistical Research Unit, Department of Economics, University of Gothenburg, Gothenburg, Sweden

Howard Burkom, Johns Hopkins Applied Physics Laboratory, Laurel, Maryland, USA

Marianne Frisén, University of Gothenburg, Gothenburg, Sweden

Webinar Logistics

ISDS
INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

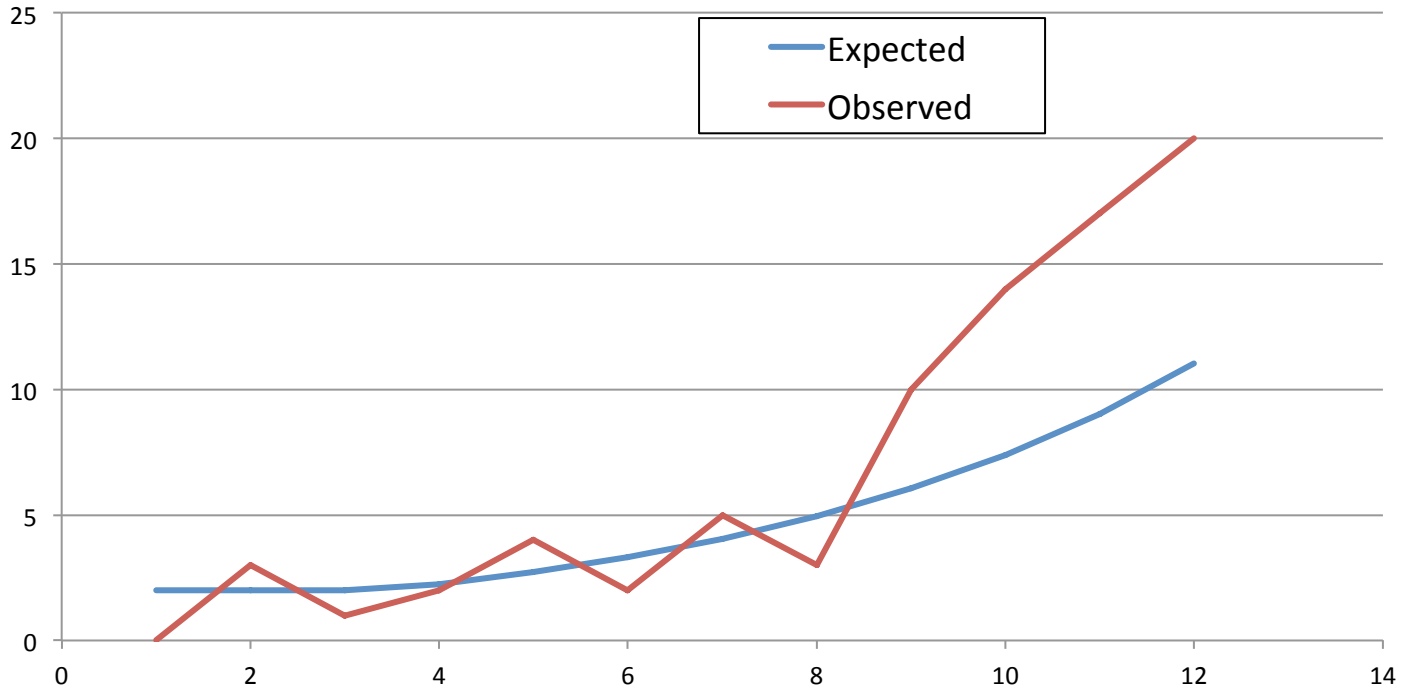
- All attendees are in listen-only mode.
- All questions will be addressed at the end of the presentation. Please feel free to type questions throughout using:
    - The question section of the GoToWebinar control panel
    - Twitter (#ProjectTychoWebinar)
- During the Q&A portion of the webinar, you may also use the GoToWebinar control panel to virtually raise your hand to be unmuted
- Evaluation: As you are leaving the webinar, you should see an evaluation pop-up. Please take a few moments to provide your input.
- 1 CPH recertification credit is available for viewing this webinar (must complete the evaluation).

What is an outbreak?

- Many different definitions
- Here are some examples:

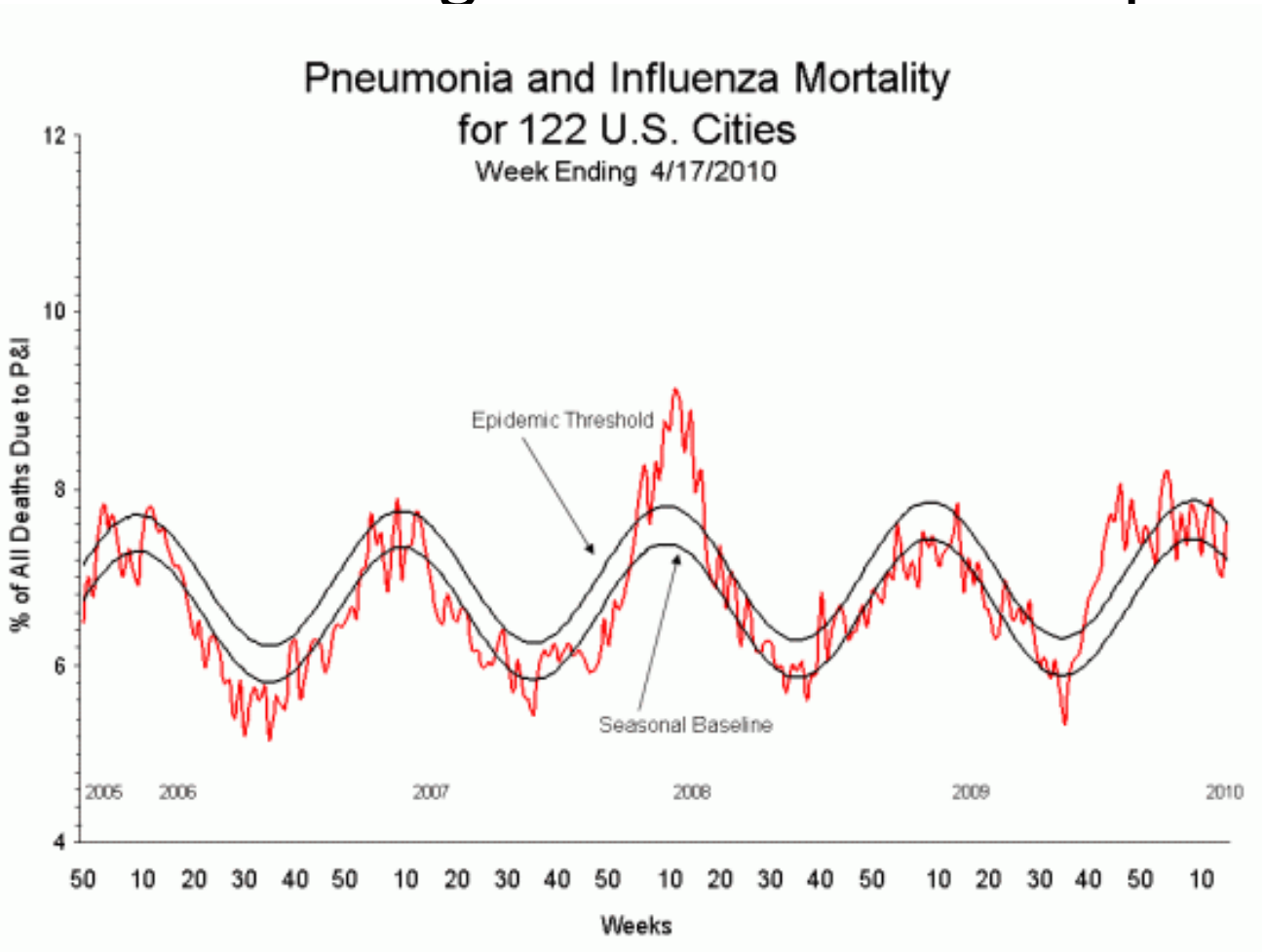
What is an outbreak?

1. An incidence higher than what is expected



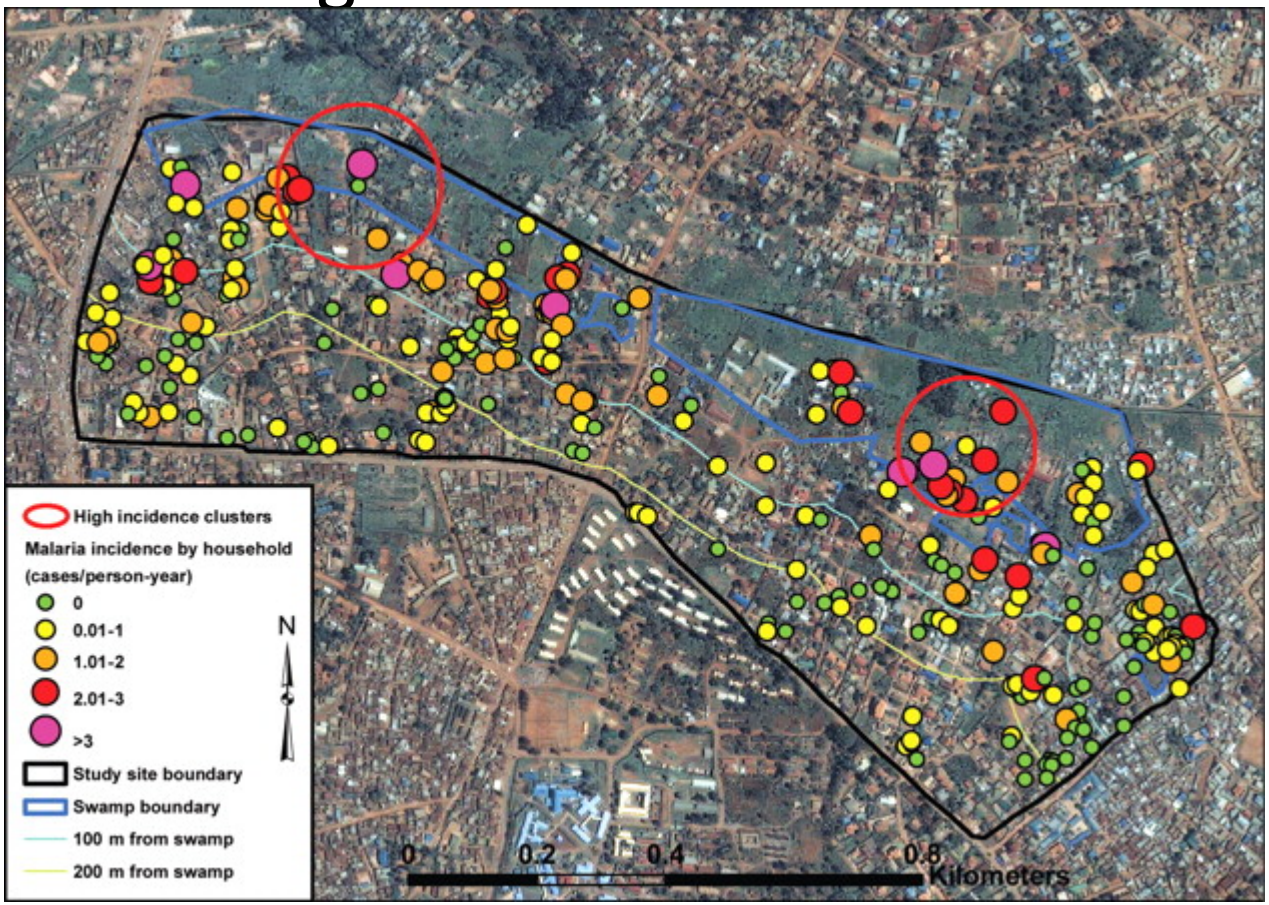
What is an outbreak?

1. An incidence higher than what is expected



What is an outbreak?

2. Clustering of events close in location and time

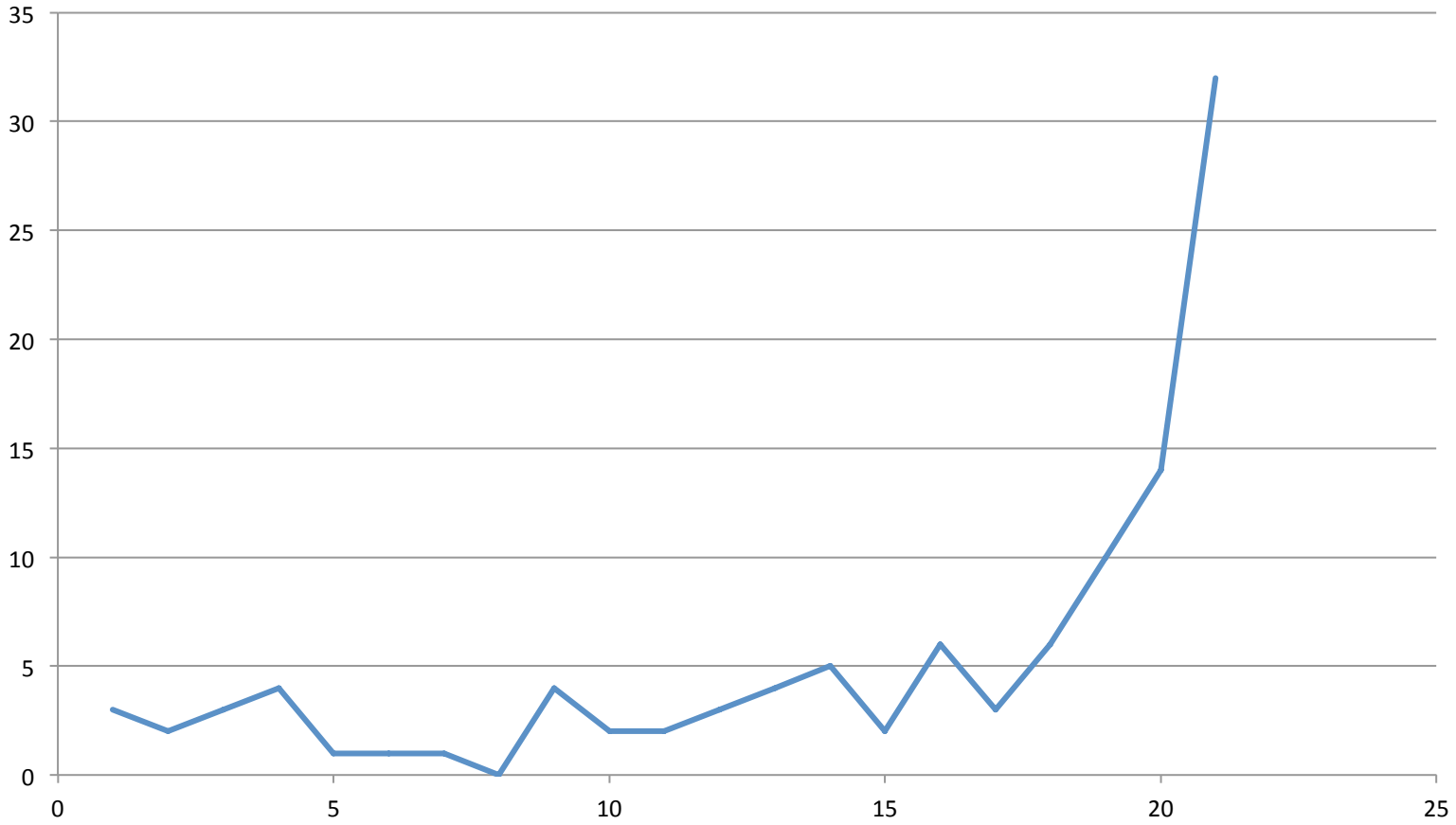


What is an outbreak?

3. A number of confirmed cases where the infection is from a local source

What is an outbreak?

4. A change from constant level to increasing

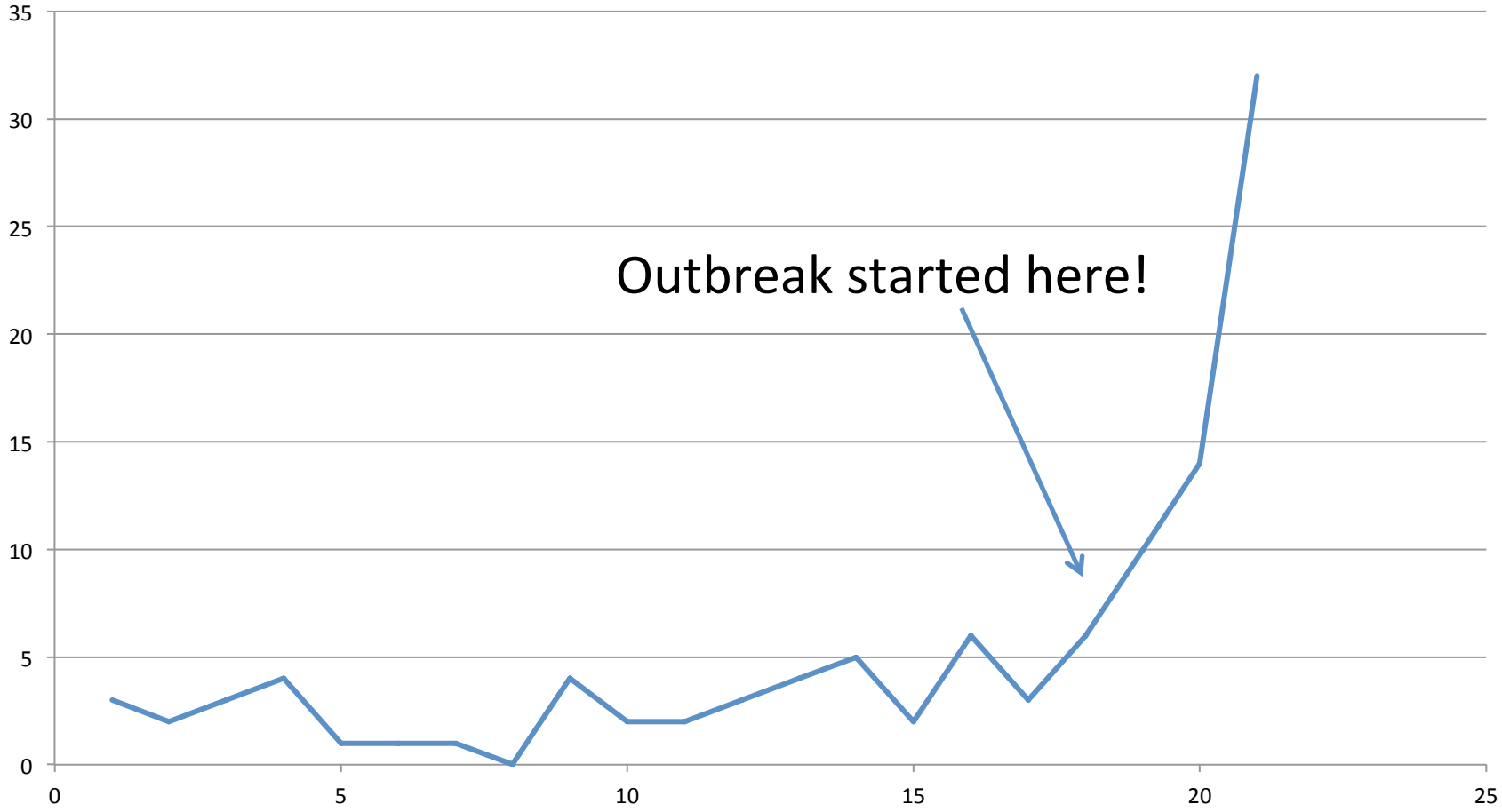


Statistical Surveillance

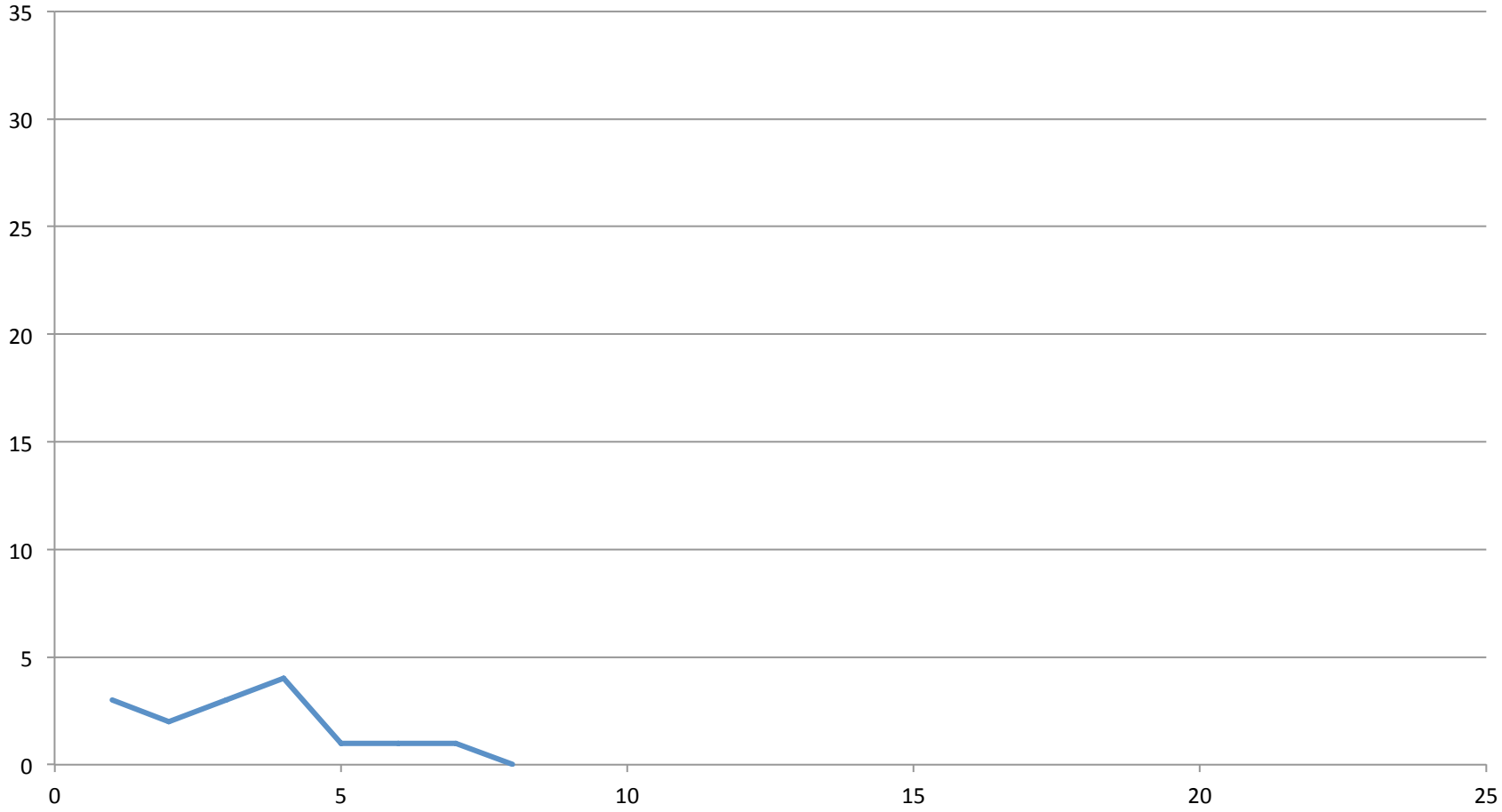
- Online detection
- Evaluate information each time a new observation is made

- Compare the following:

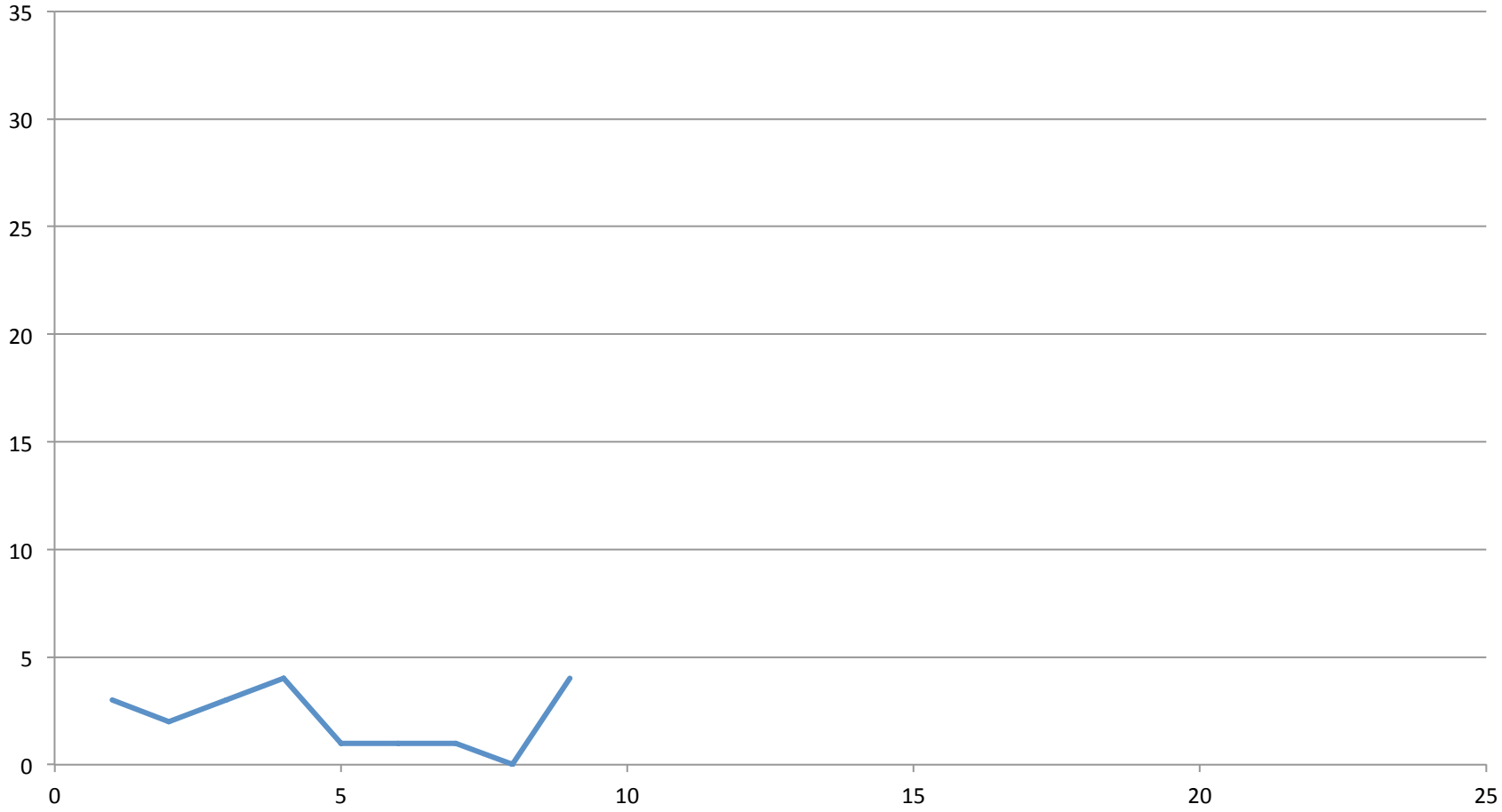
Statistical Surveillance



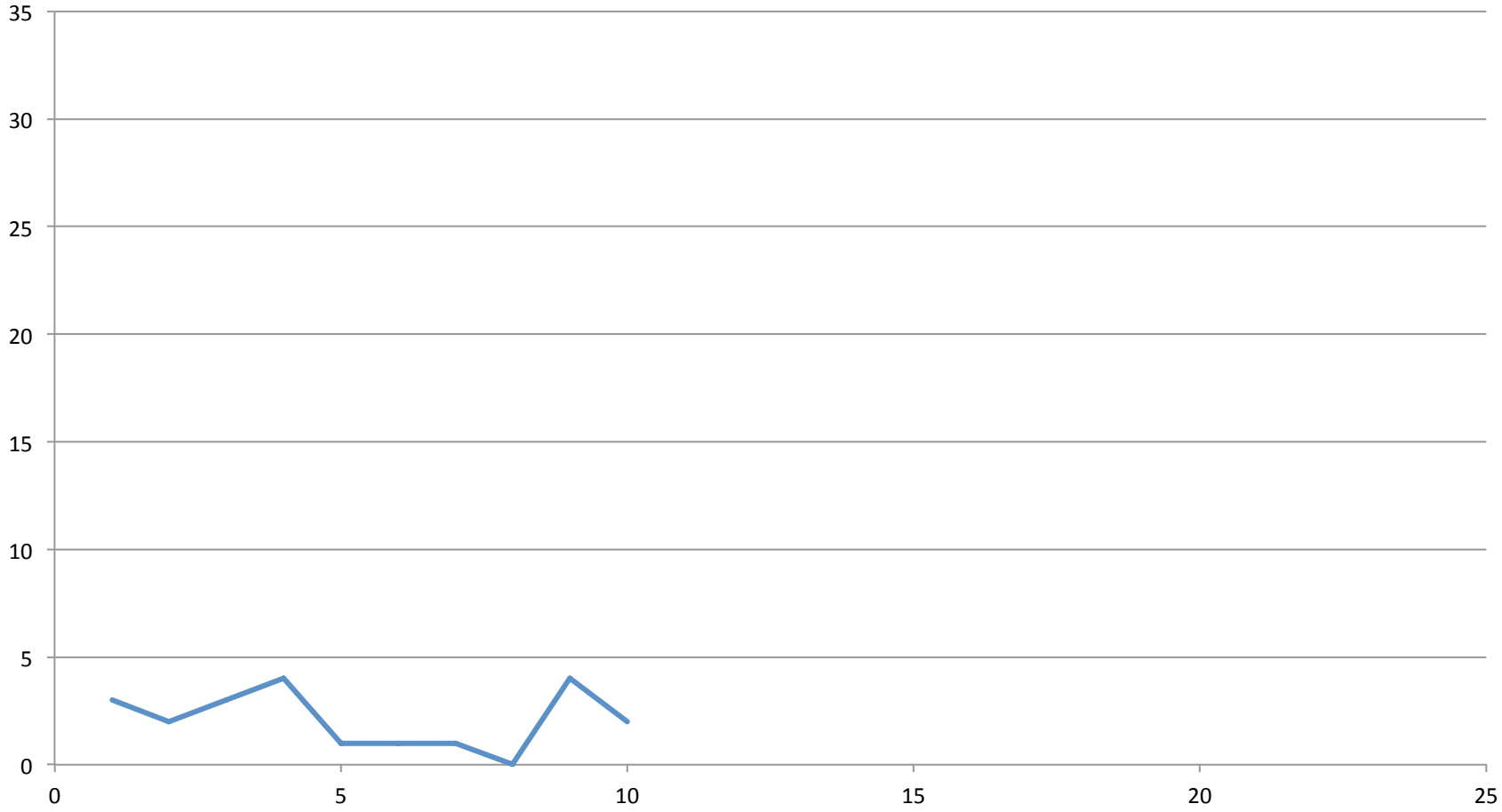
Statistical Surveillance



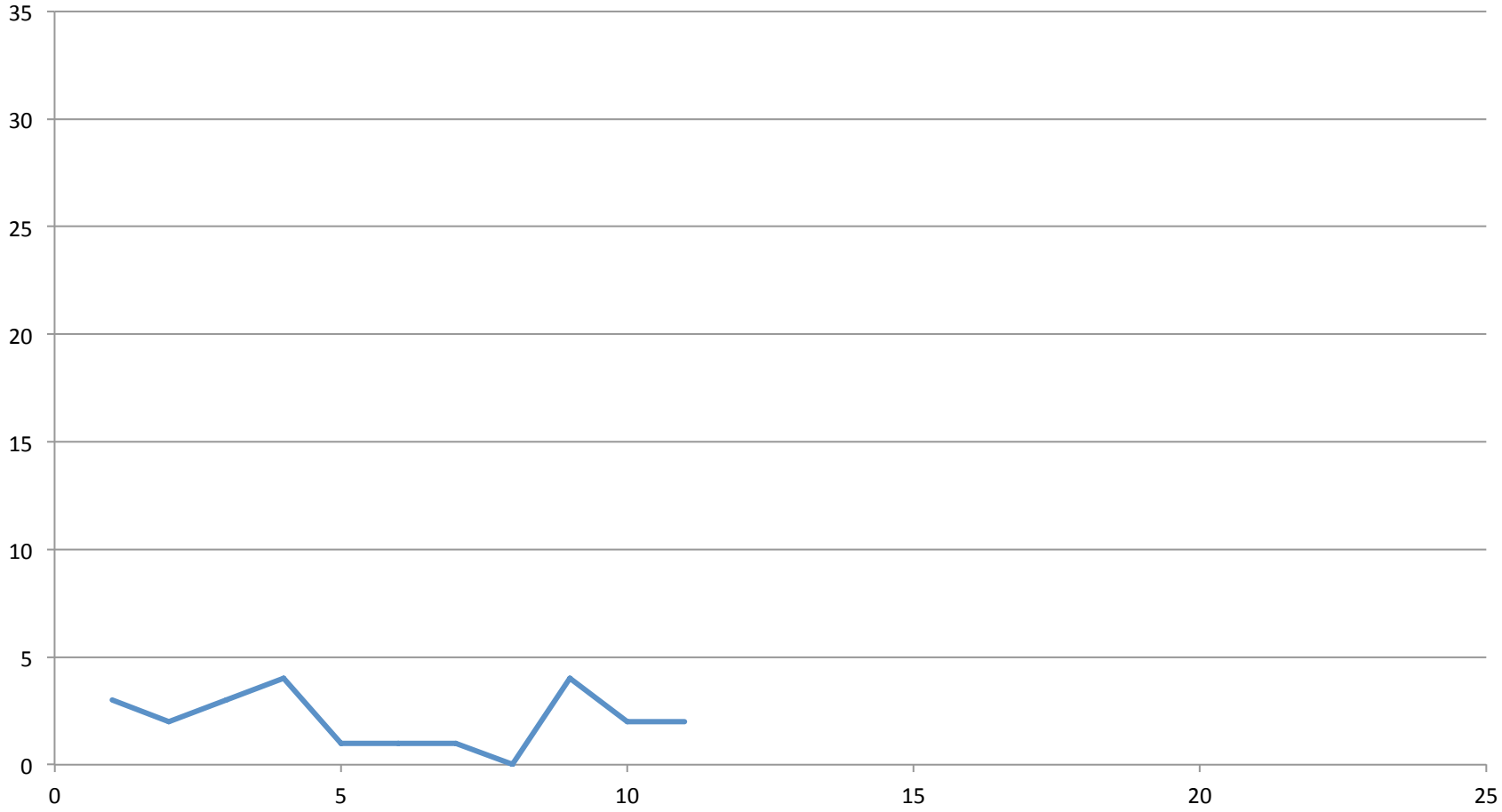
Statistical Surveillance



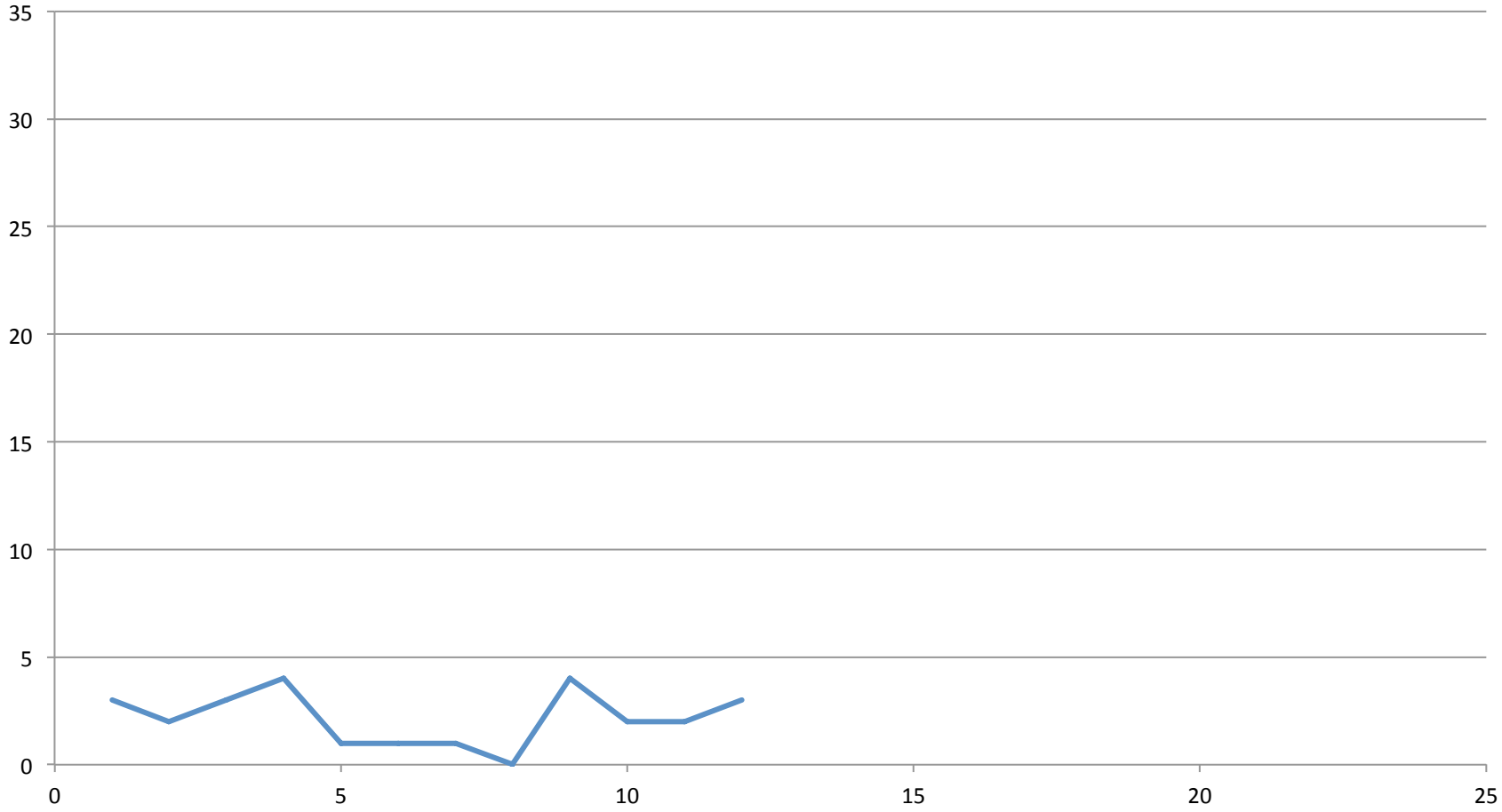
Statistical Surveillance



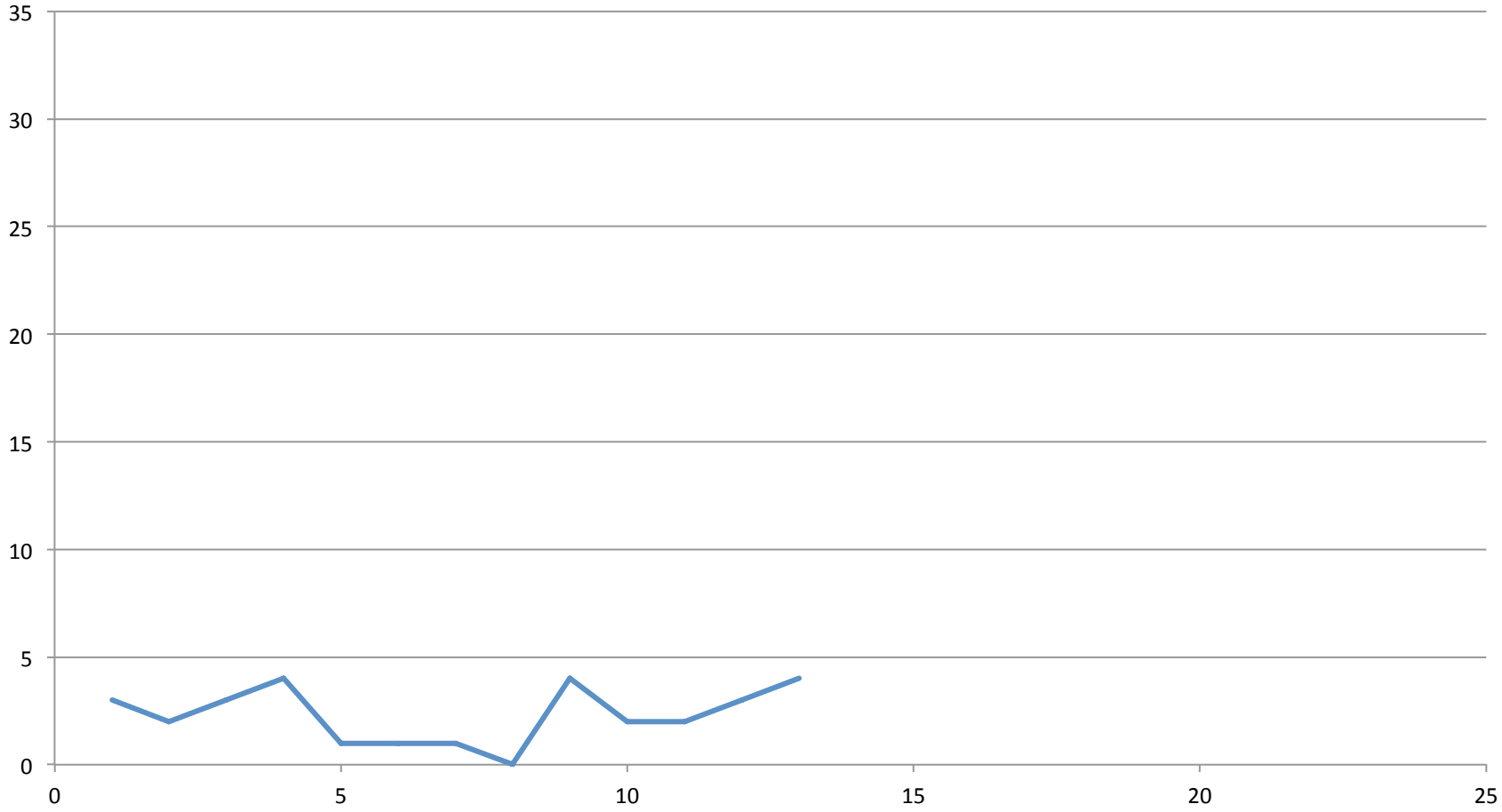
Statistical Surveillance



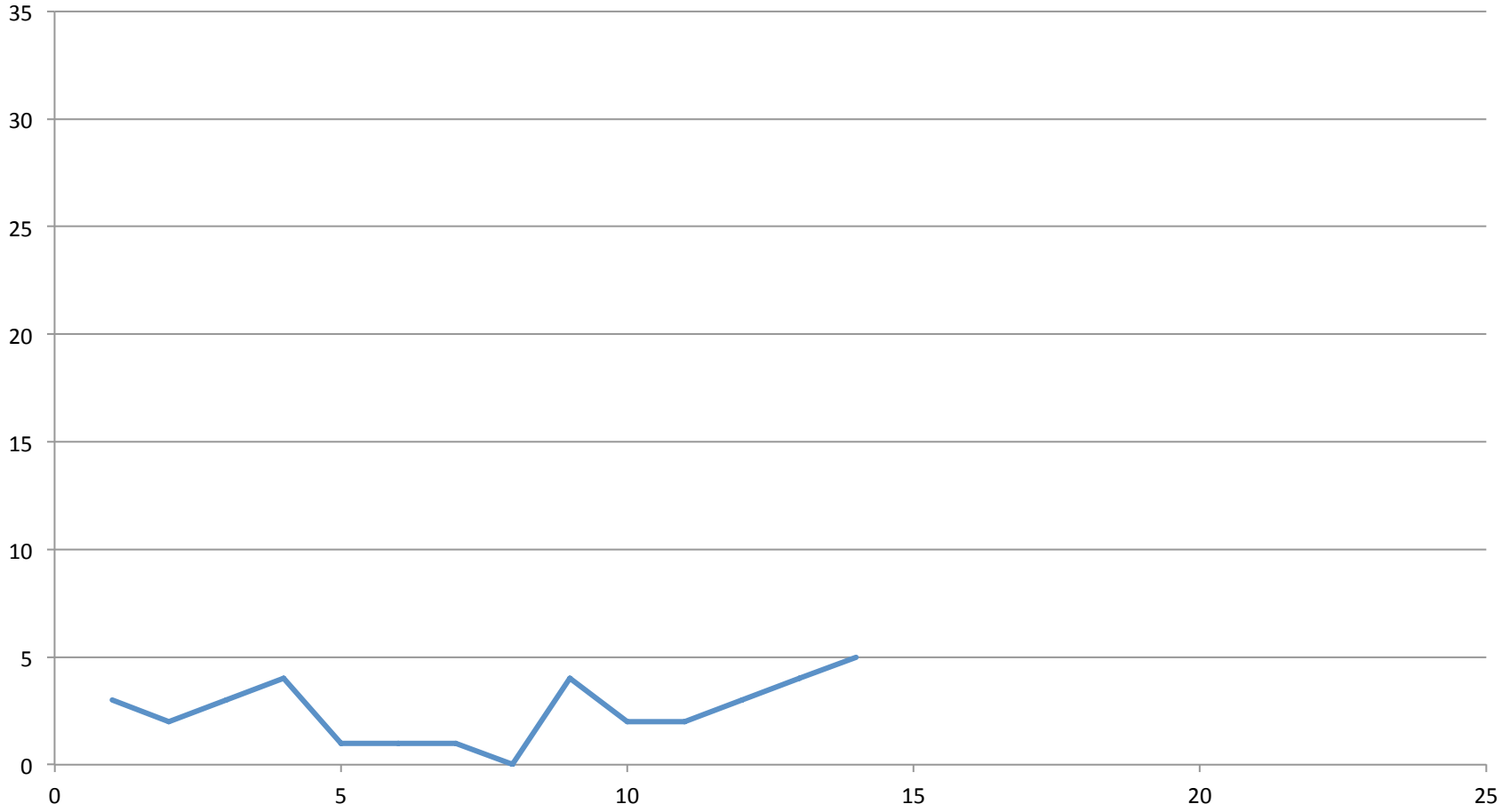
Statistical Surveillance



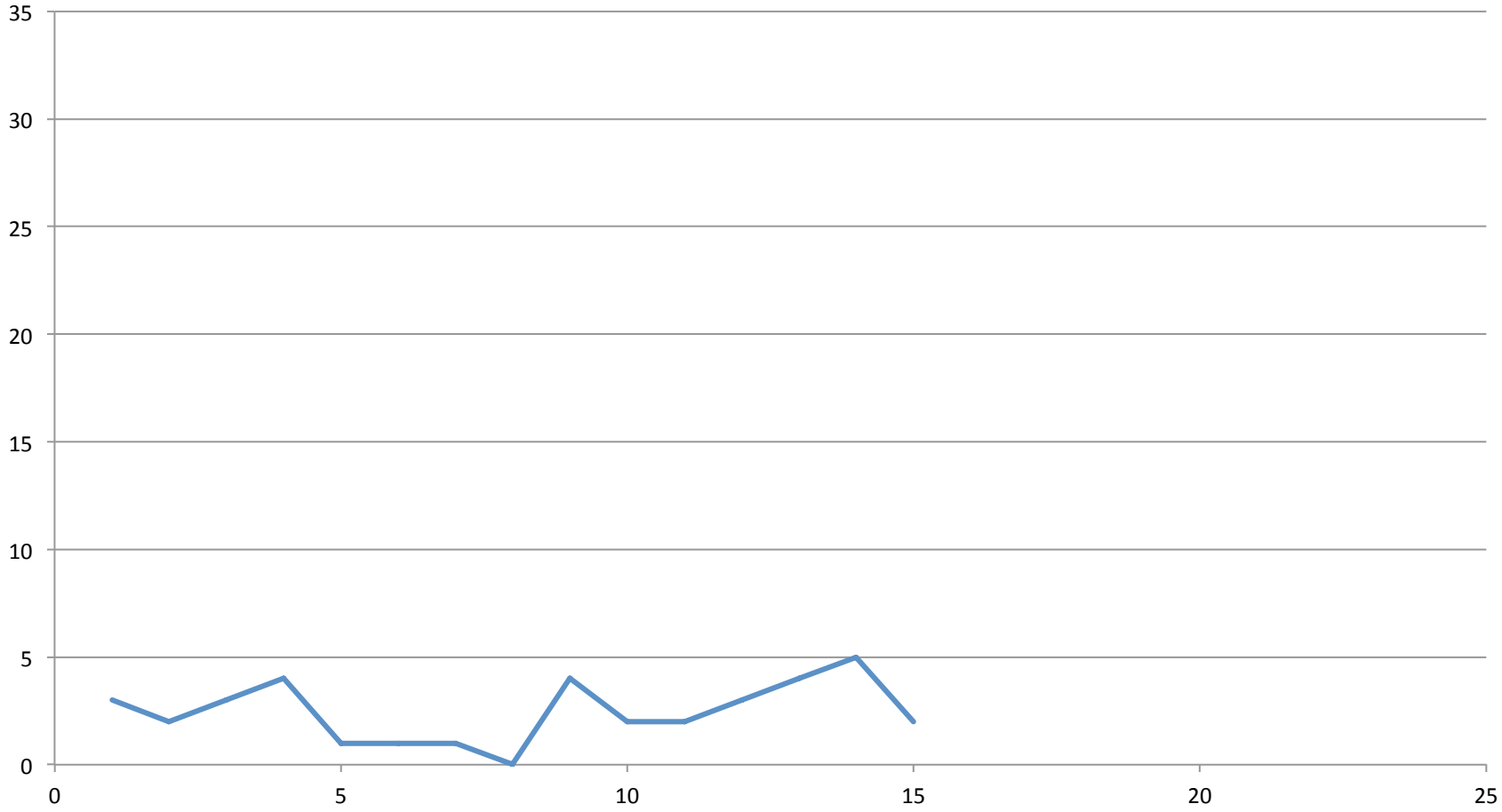
Statistical Surveillance



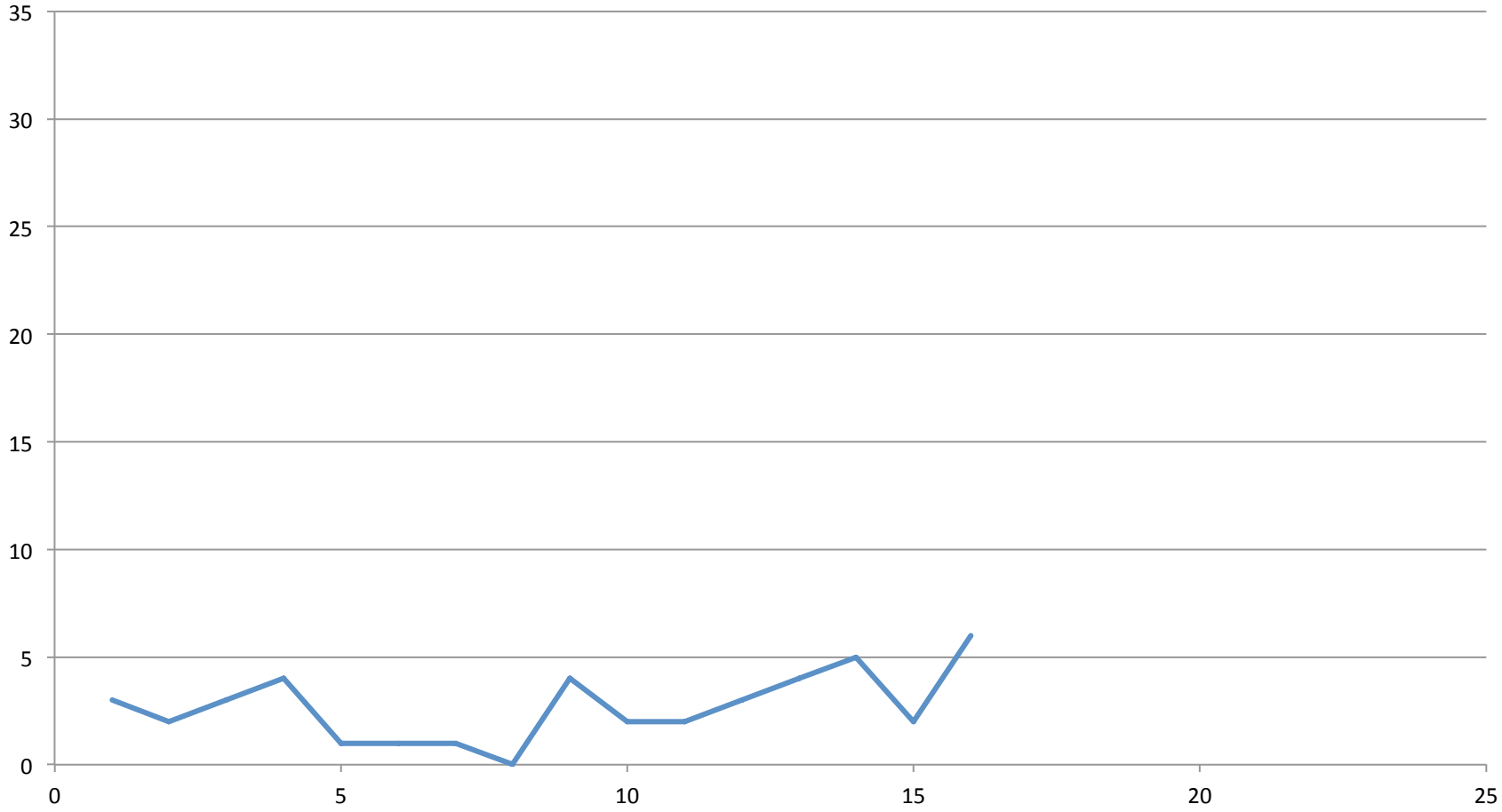
Statistical Surveillance



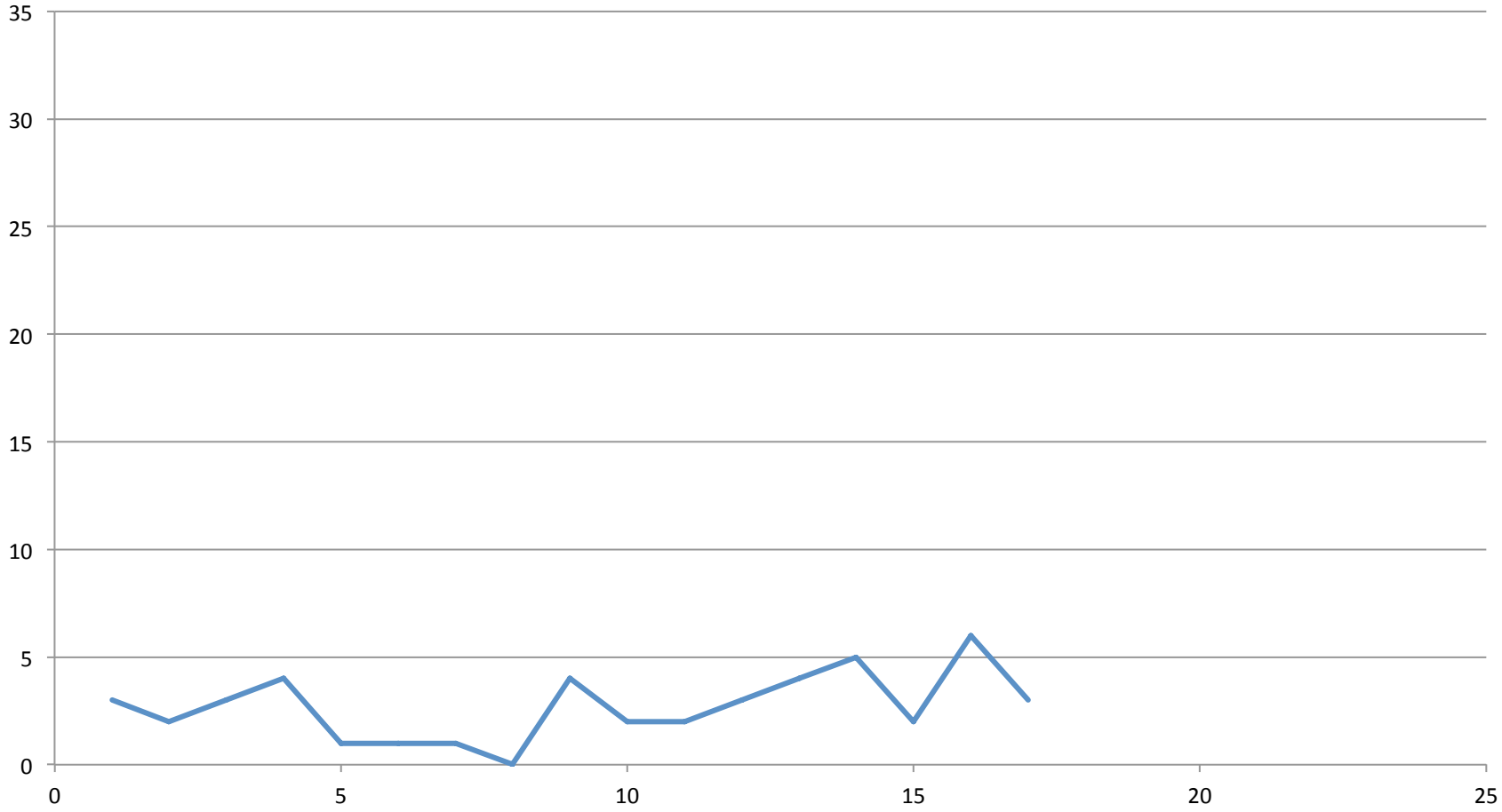
Statistical Surveillance



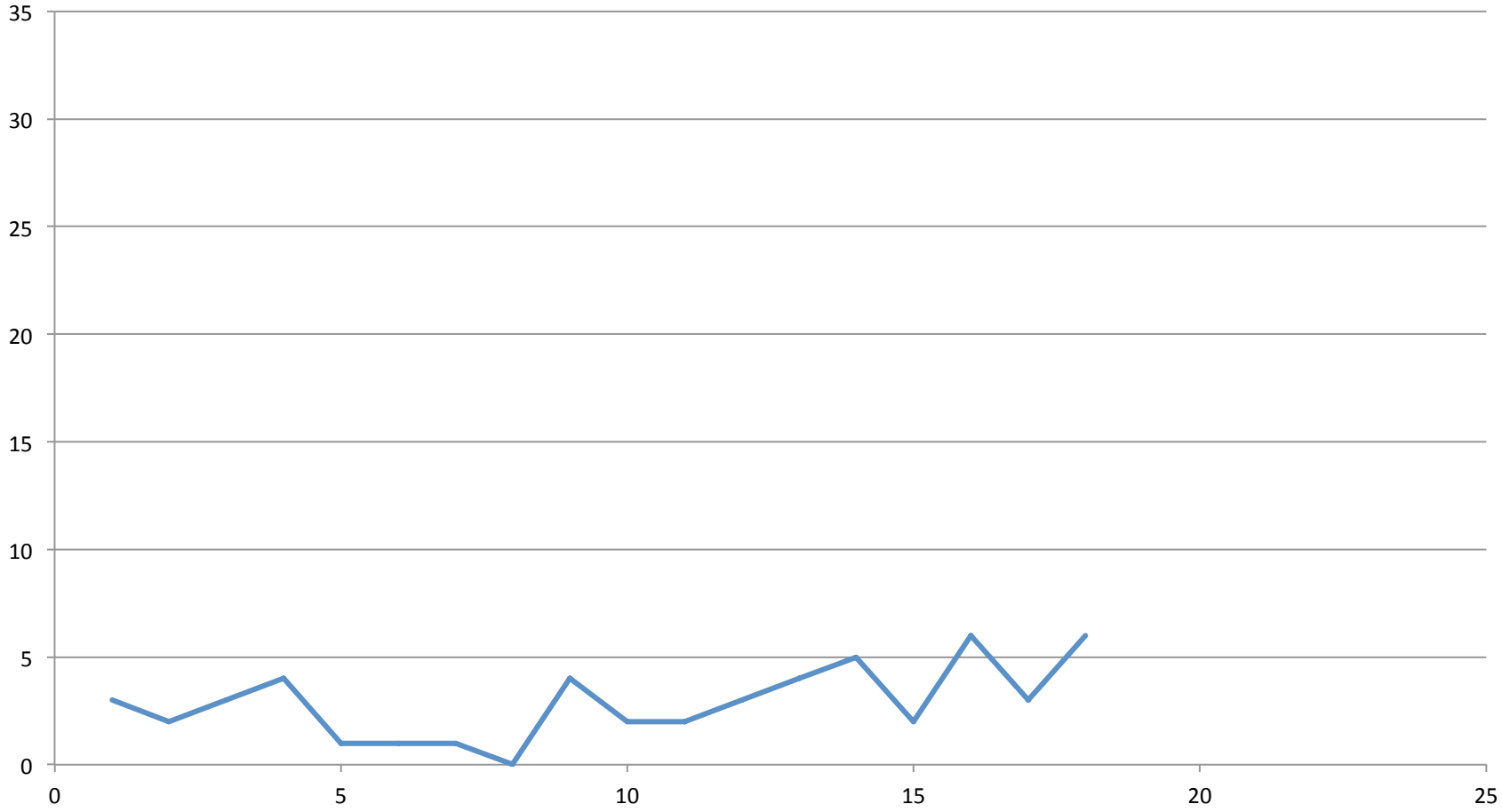
Statistical Surveillance



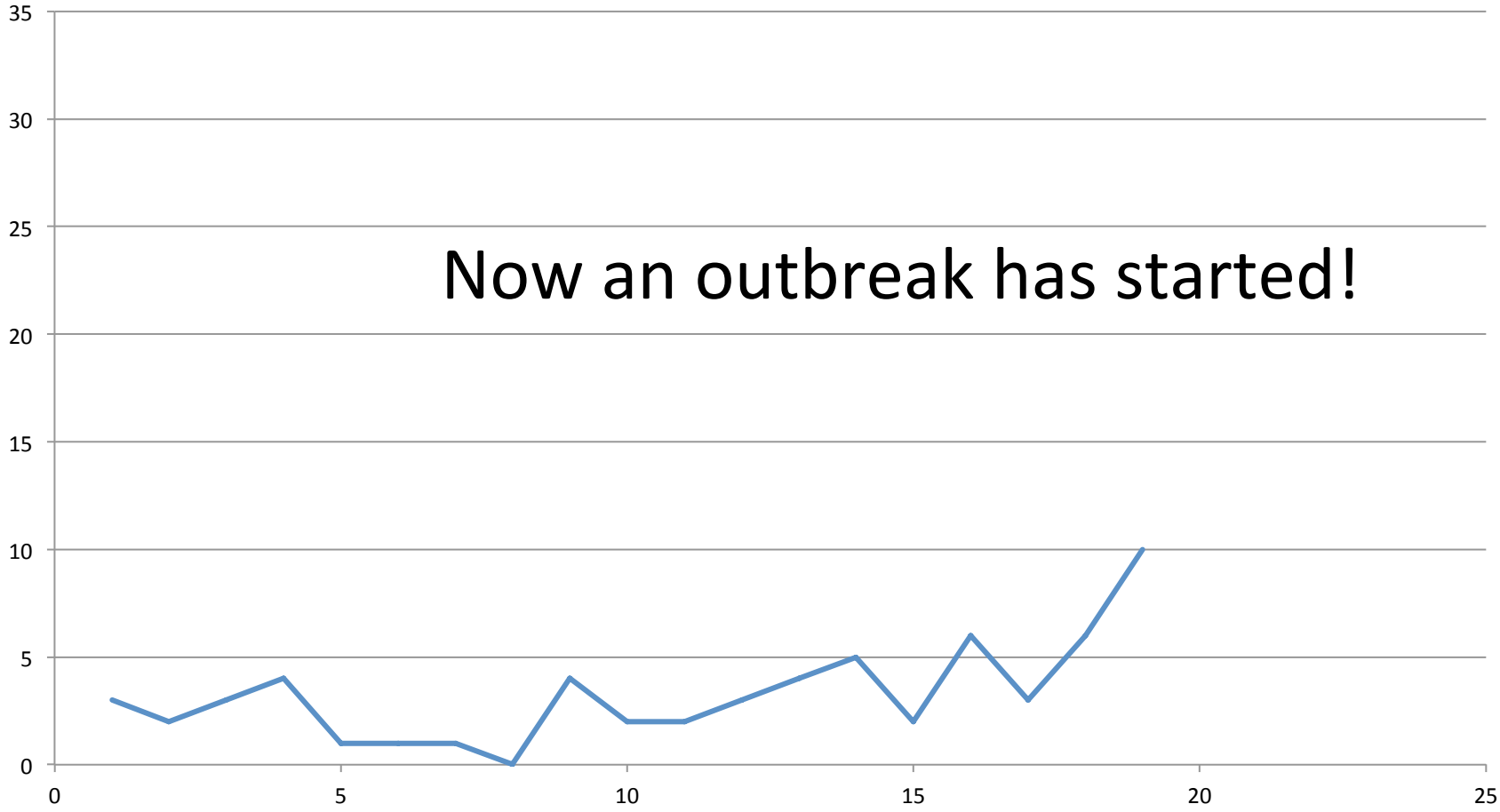
Statistical Surveillance



Statistical Surveillance



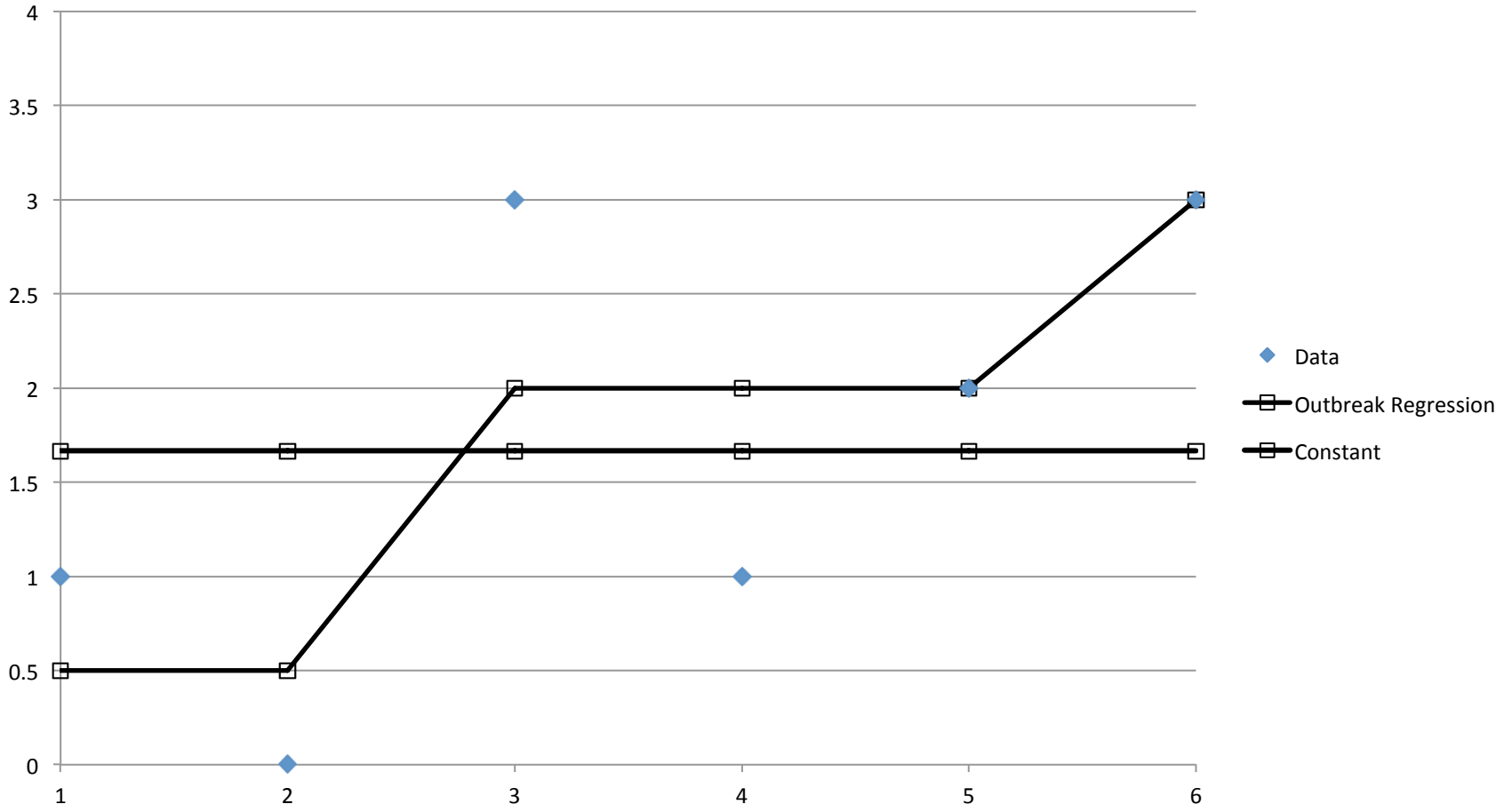
Statistical Surveillance



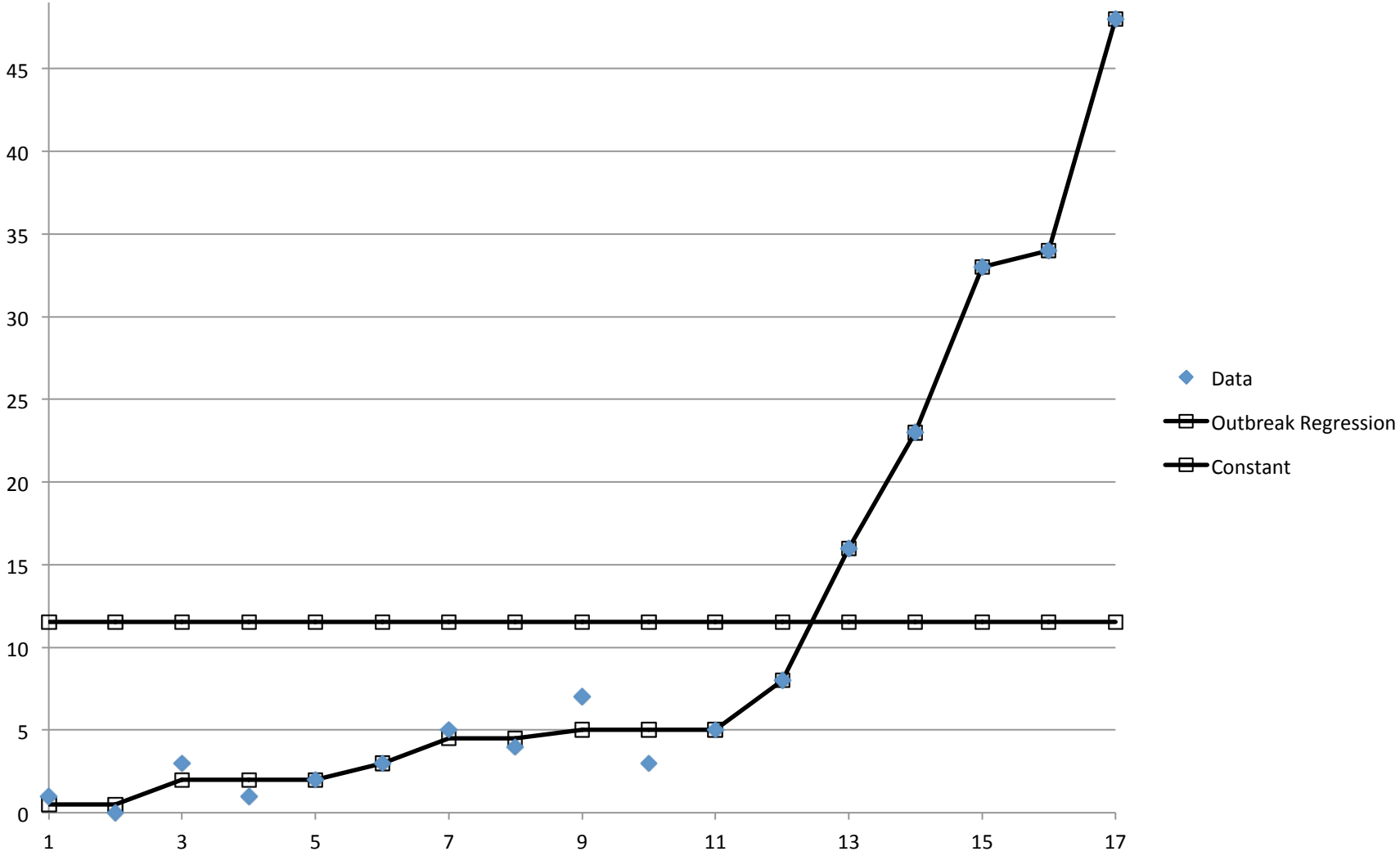
OutbreakP

- We developed a semiparametric method for outbreak detection
- Relies on the shape of the curve
- Likelihood ratio between restriction of increasing level and constant level

OutbreakP



OutbreakP


45
40
35
30
25
20
15
10
5
0
1
3
5
7
9
11
13
15
17
Data
Outbreak Regression
Constant

Measures of performance

- We want a method that
  - Detects outbreaks without small delays
  - Gives few false alarms
- Balance between few false alarm and delay
  - Faster detection -> more false alarms
  - More accurate detection -> longer delay
- We need measures to evaluate this!
- Multivariate surveillance needs special measures
- Frisén, M., E. Andersson, et al. (2010). "Evaluation of Multivariate Surveillance." Journal of Applied Statistics **37(12): 2089-2100.**

Measures of performance

- Notation:
  - t_A - time of alarm
  - τ - time of outbreak
- Conditional Expected Delay, CED

$$CED(\tau) = E\left[t_A - \tau \middle| t_A \geq \tau\right]$$

- Average Run Length

$$ARL^0 = E\left[t_A \mid \tau = \infty\right]$$

- Skewed distribution
  - Median Run Length0, MRL^0

Measures of performance

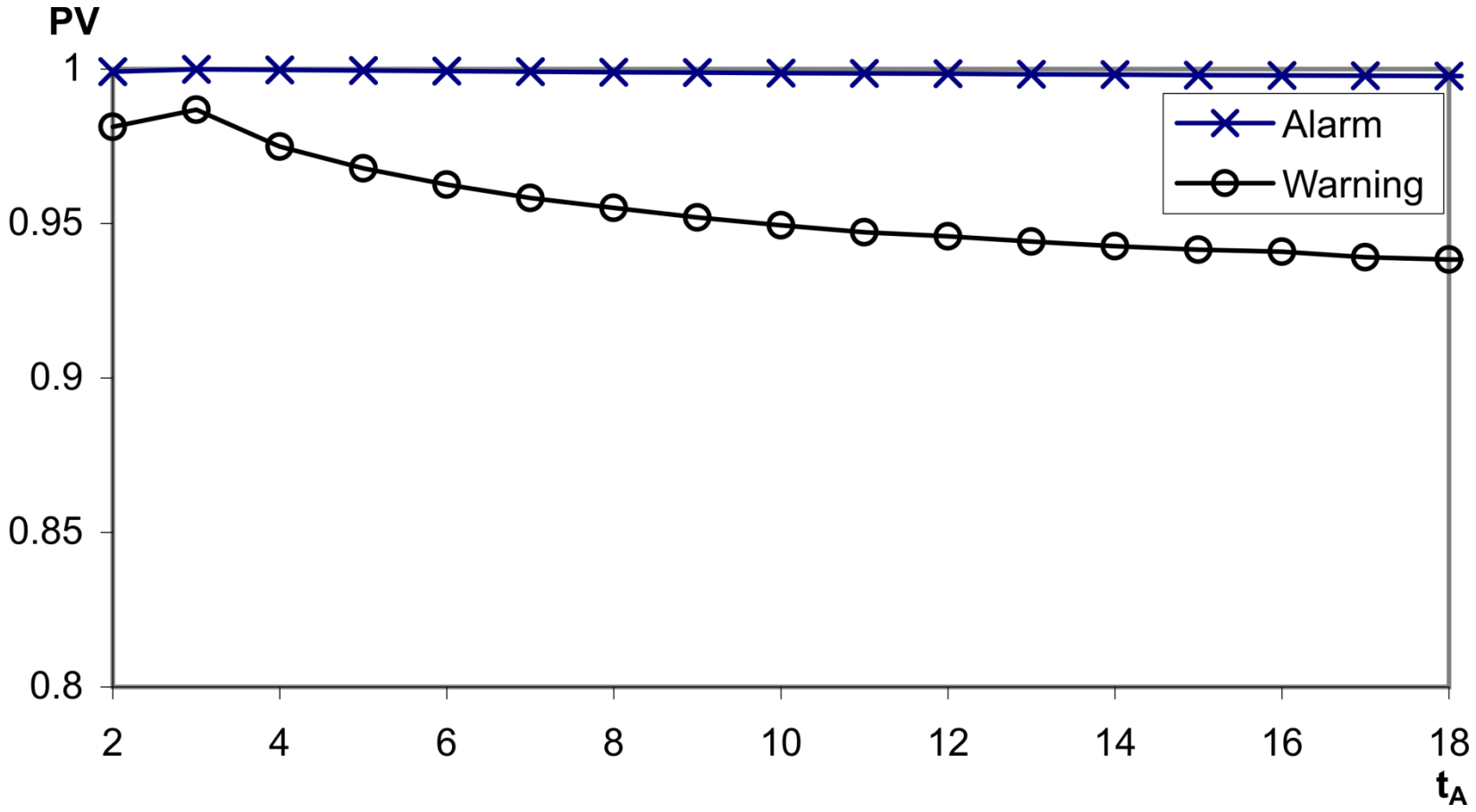
- Predictive Value, PV

$$PV(t) = P(\tau \le t \mid t_A = t) = \frac{\sum_{i=1}^{t} \left(P(t_A = t \mid \tau = i) P(\tau = i) \right)}{\sum_{i=1}^{t} \left(P(t_A = t \mid \tau = i) P(\tau = i) \right) + P(t_A = t \mid \tau > t) P(\tau > t)}$$

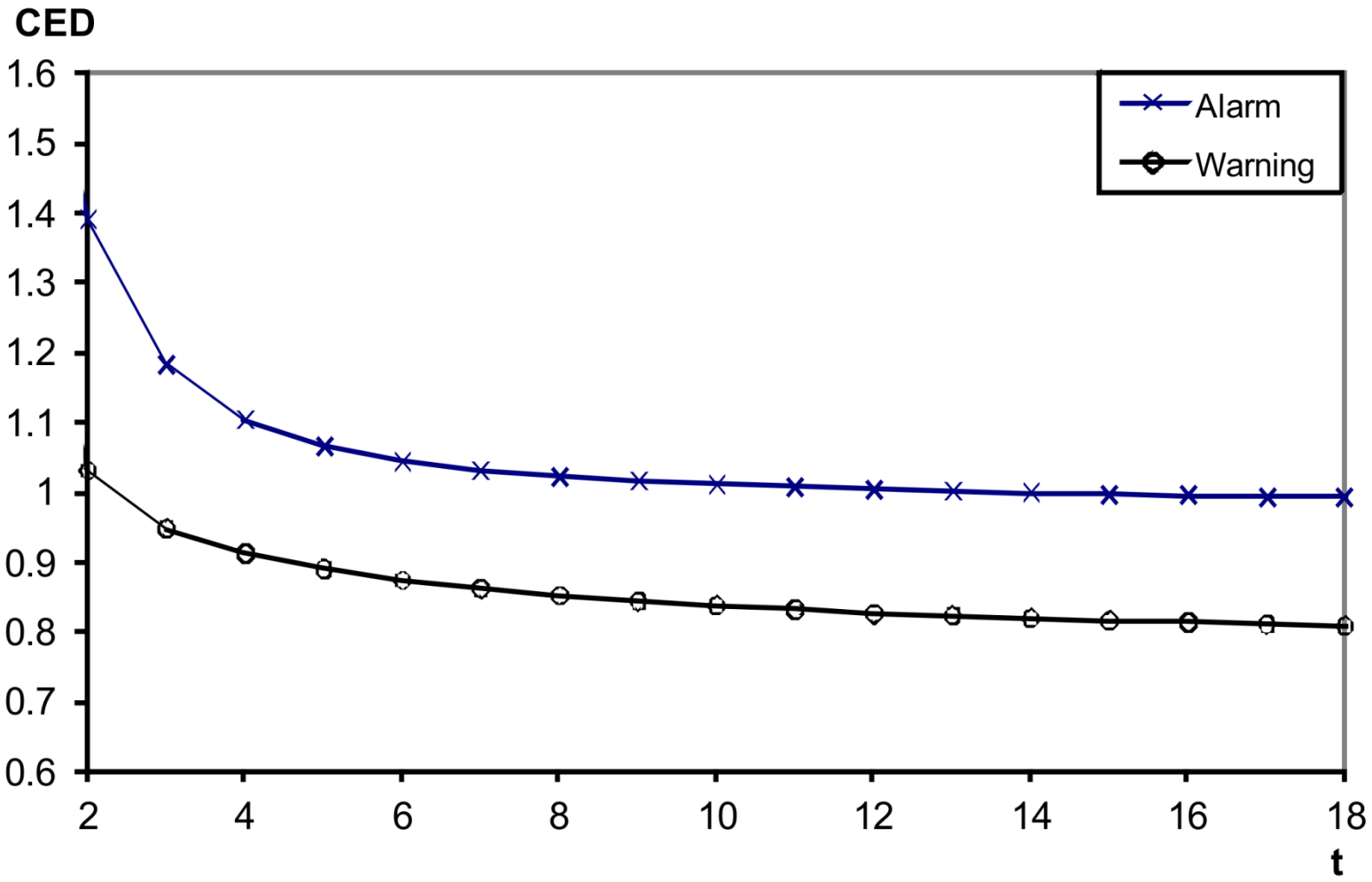
- Or

$$p\left(\text{alarm at t is true}\right) = \frac{p\left(\text{true alarm at/before t}\right)}{p\left(\text{true alarm at/before t}\right) + p\left(\text{false alarm after t}\right)}$$

Predictive Value



Conditional Expected Delay



Multivariate surveillance

- Monitoring of p processes instead of one
- What to detect?
  - Time of first change in any process
  - Time of change in a specific process
  - Time of change in at least n processes

Updated Measures of performance

- Notation:
  - t_A - time of alarm
  - $\tau_{\min}$ - time of first outbreak
- Conditional Expected Delay, CED

$$CED\left(\tau_1,\ \tau_2,\ \ldots,\ \tau_p\right) = E\left[t_A - \tau_{\min} \,\middle|\, t_A \geq \tau_{\min}\right]$$

- Predictive Value, PV

$$PV(t) = P(\tau_{\min} \leq t \mid t_A = t) = \frac{\sum_{i=1}^{t}\left(P(t_A = t \mid \tau_{\min} = i)P(\tau_{\min} = i)\right)}{\sum_{i=1}^{t}\left(P(t_A = t \mid \tau_{\min} = i)P(\tau_{\min} = i)\right) + P(t_A = t \mid \tau_{\min} > t)P(\tau_{\min} > t)}$$

Multivariate surveillance

- Different approaches:
  - Parallel univariate surveillance
  - Reduction of dimension
    - Reduction to a scalar
  - Vector Accumulation
  - Using full multivariate likelihood
    - Can be quite complex

The sufficiency principle

- A sufficient statistic captures all the information about a parameter θ available in a sample
- Inference regarding θ should be the same for the same value of an observed statistic $T(\mathbf{X})=t(\mathbf{x})$
- Assumptions important!

Multivariate surveillance

- Data reduction to reduce complexity
- Sufficient reduction – no information is lost
- We showed that the sum is sufficient for surveillance in the one parameter exponential family
- The semiparametric method further developed

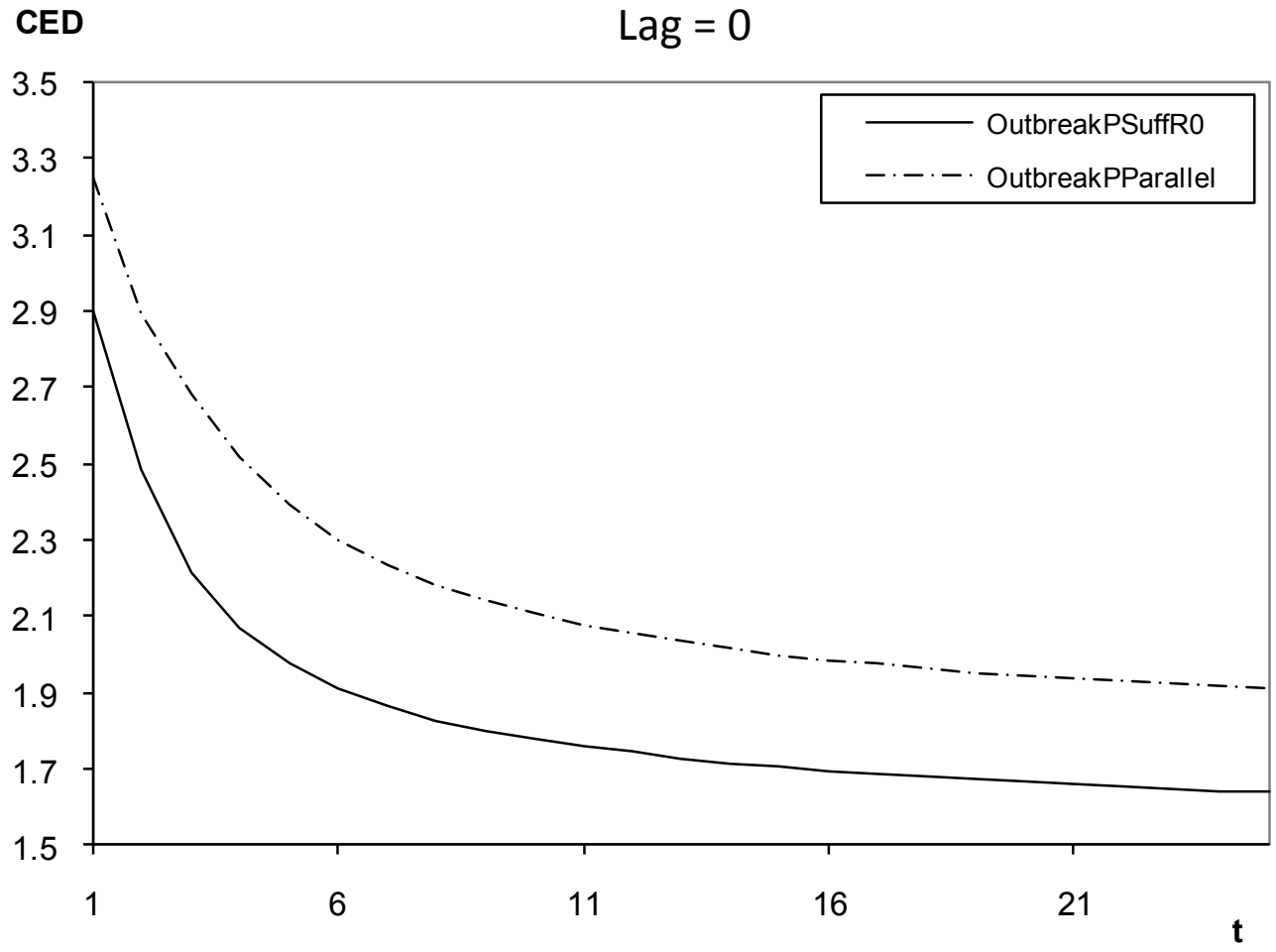
Application of the sufficiency principle to Statistical surveillance

- Known time-lag between the processes
- Processes iid given time and lag
  - Same baseline, same shape
- Sufficient reduction into a univariate statistic may be found

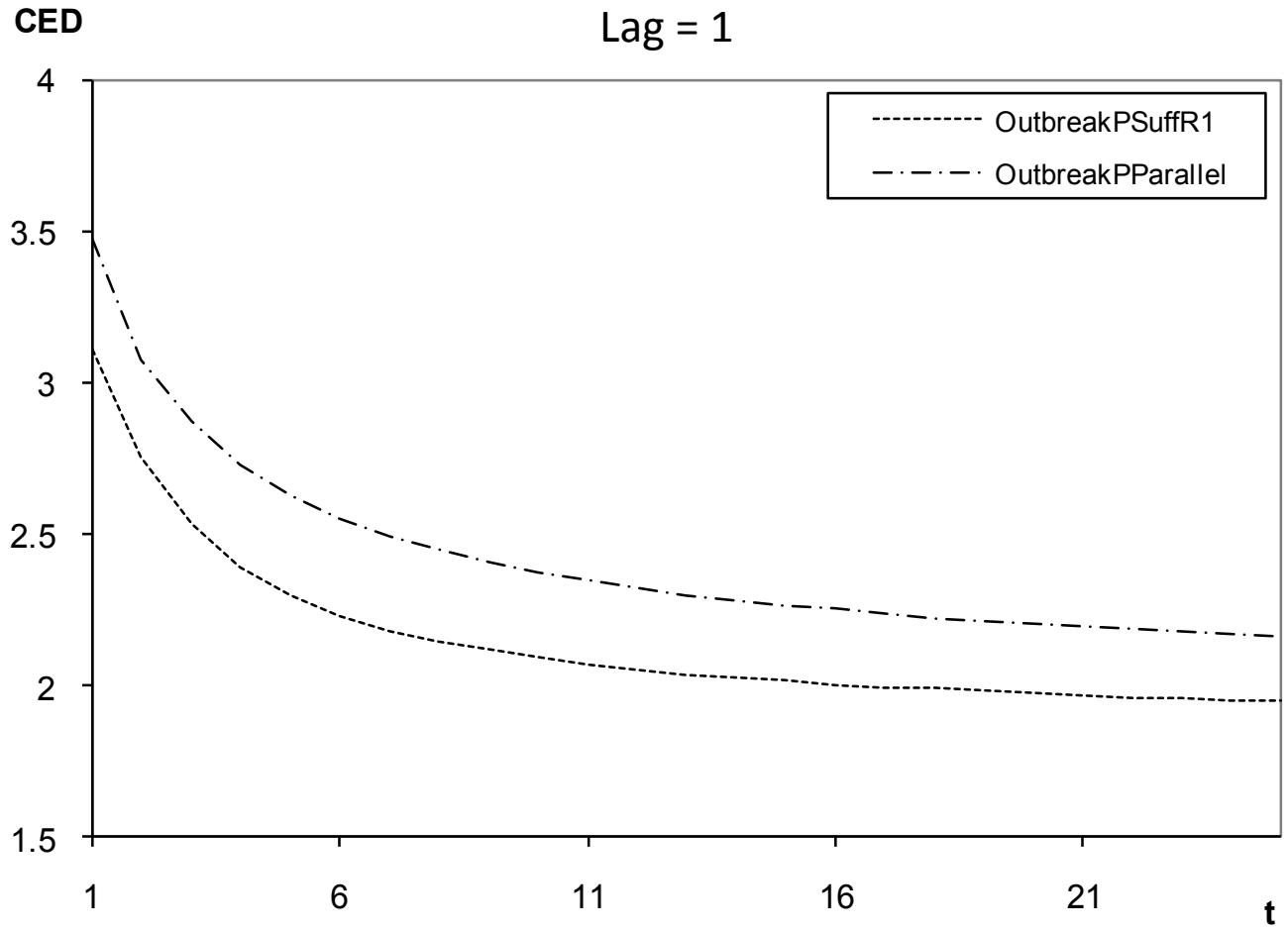
Application of the sufficiency principle to Statistical surveillance

- For p independent processes in the one-parameter exponential family a sufficient reduction is shown for step changes in
  - Frisén, M., E. Andersson, et al. (2011). "Sufficient reduction in multivariate surveillance." Communications in Statistics - Theory and Methods **40(10): 1821-1838.**
- For gradual (i.e. non-decreasing) changes
  - Schiöler, L. and M. Frisén (2012). "Multivariate outbreak detection." Journal of Applied Statistics **39(2): 223-242.**

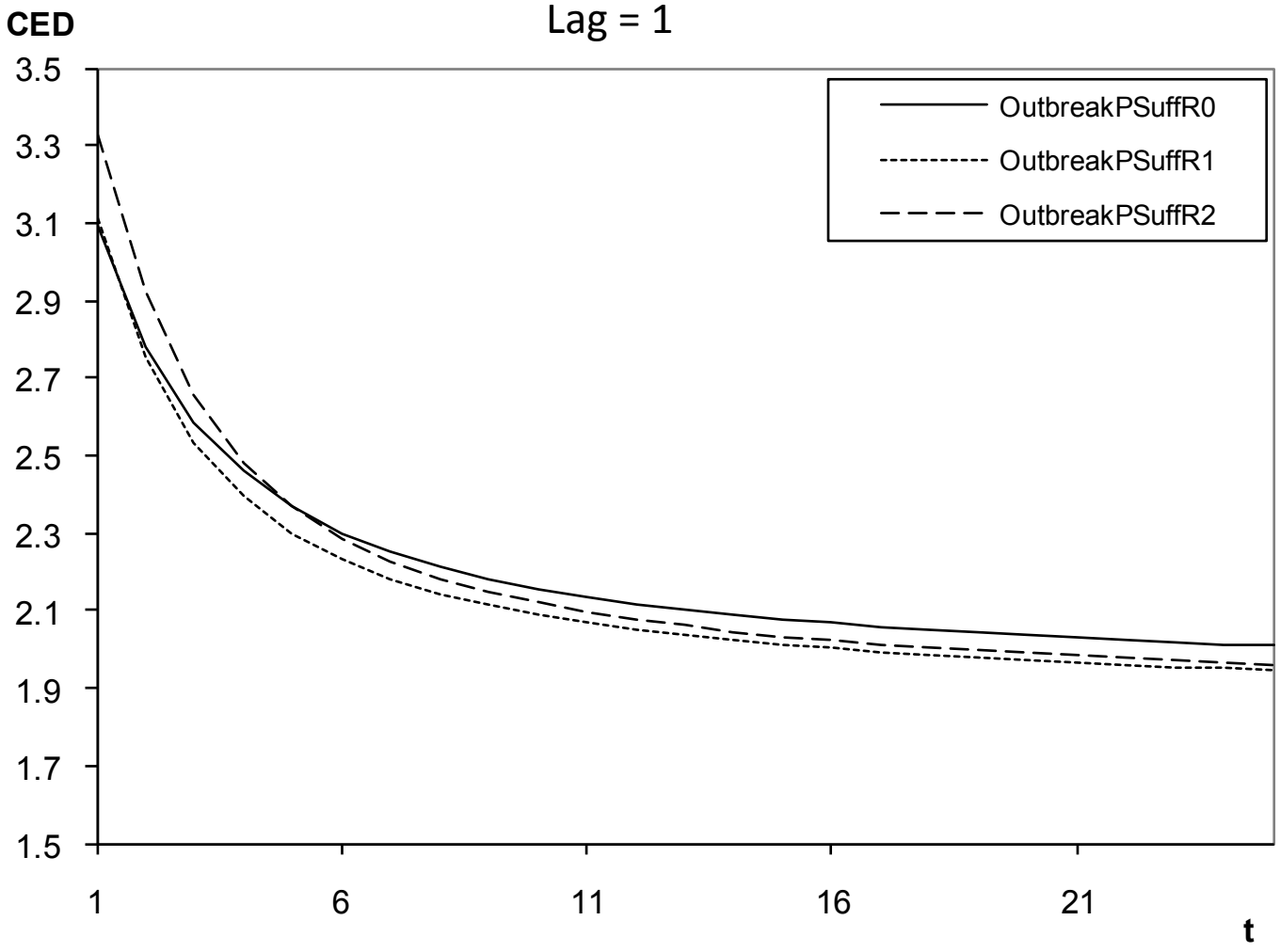
Parallel vs Reduction



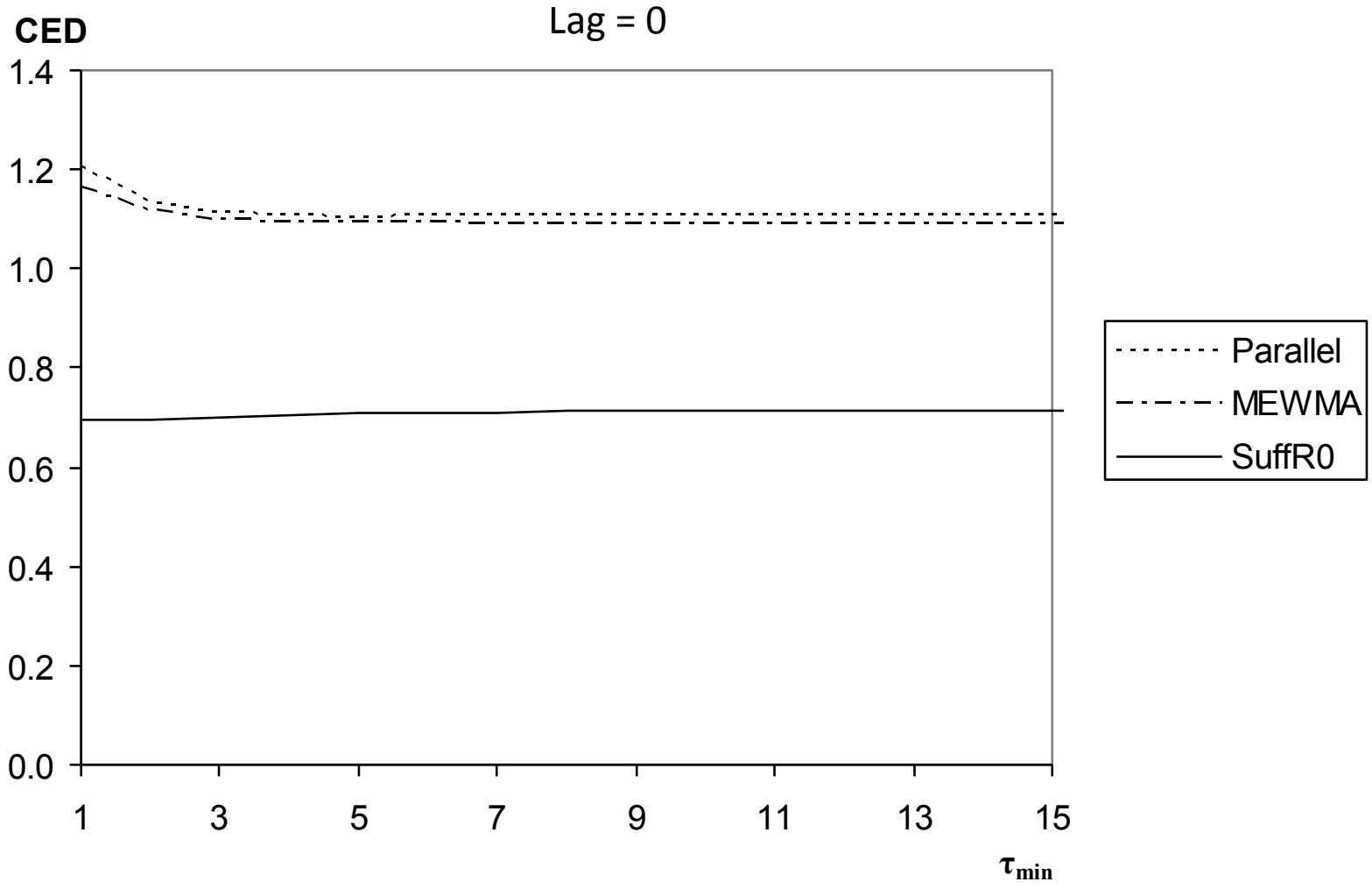
Parallel vs Reduction



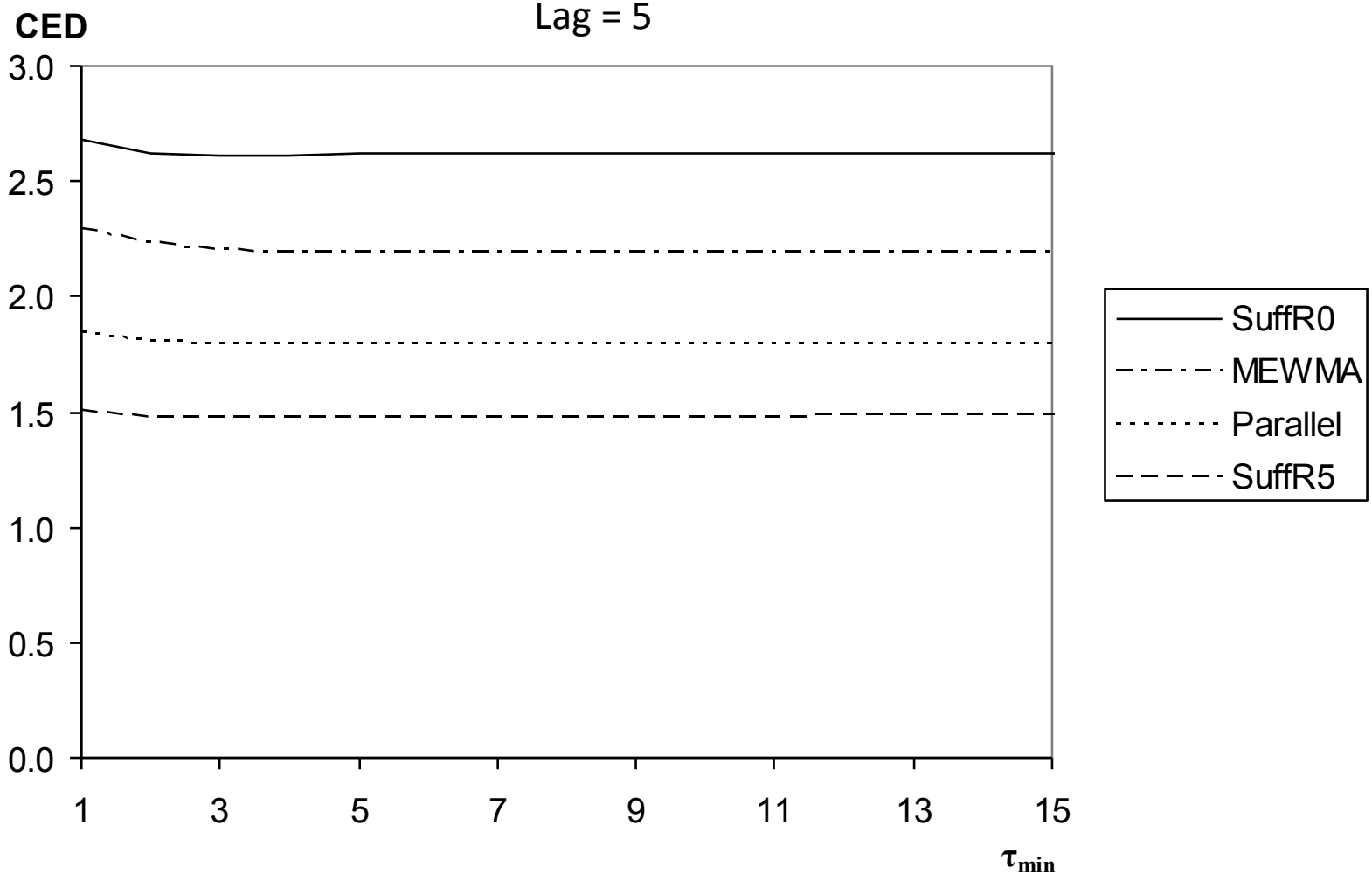
Parallel vs Reduction



Parallel vs Reduction



Parallel vs Reduction



Influenza data from Sweden

- Laboratory diagnosed influenza (LDI)
  - Collected weekly from different laboratories
  - Information on catchment area not available
  - A city may not use same laboratory each year
  - Different policies regarding testing

Data quality

- Laboratory diagnosed influenza data:
  - Baseline uncertain
  - Small number of cases, aggregation needed
  - Starting time of outbreak differs
  - Different severity of outbreak

Data quality

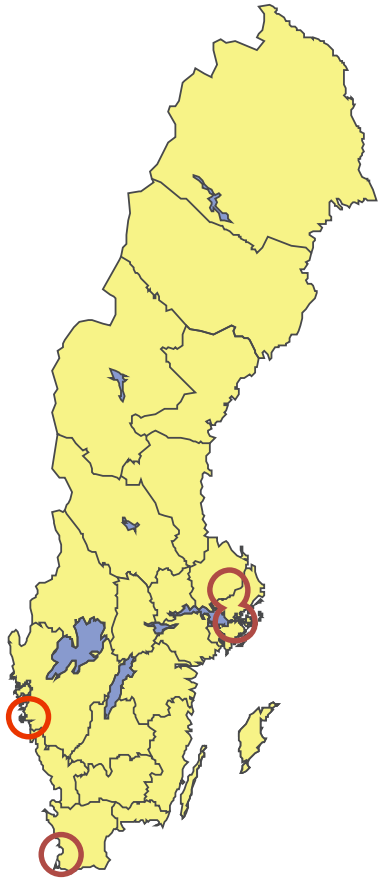
- However:
  - Data for all years from 50% of laboratories
  - Consistent reporting from large laboratories
  - Laboratories *relatively* evenly distributed with regards to population

Spatial information

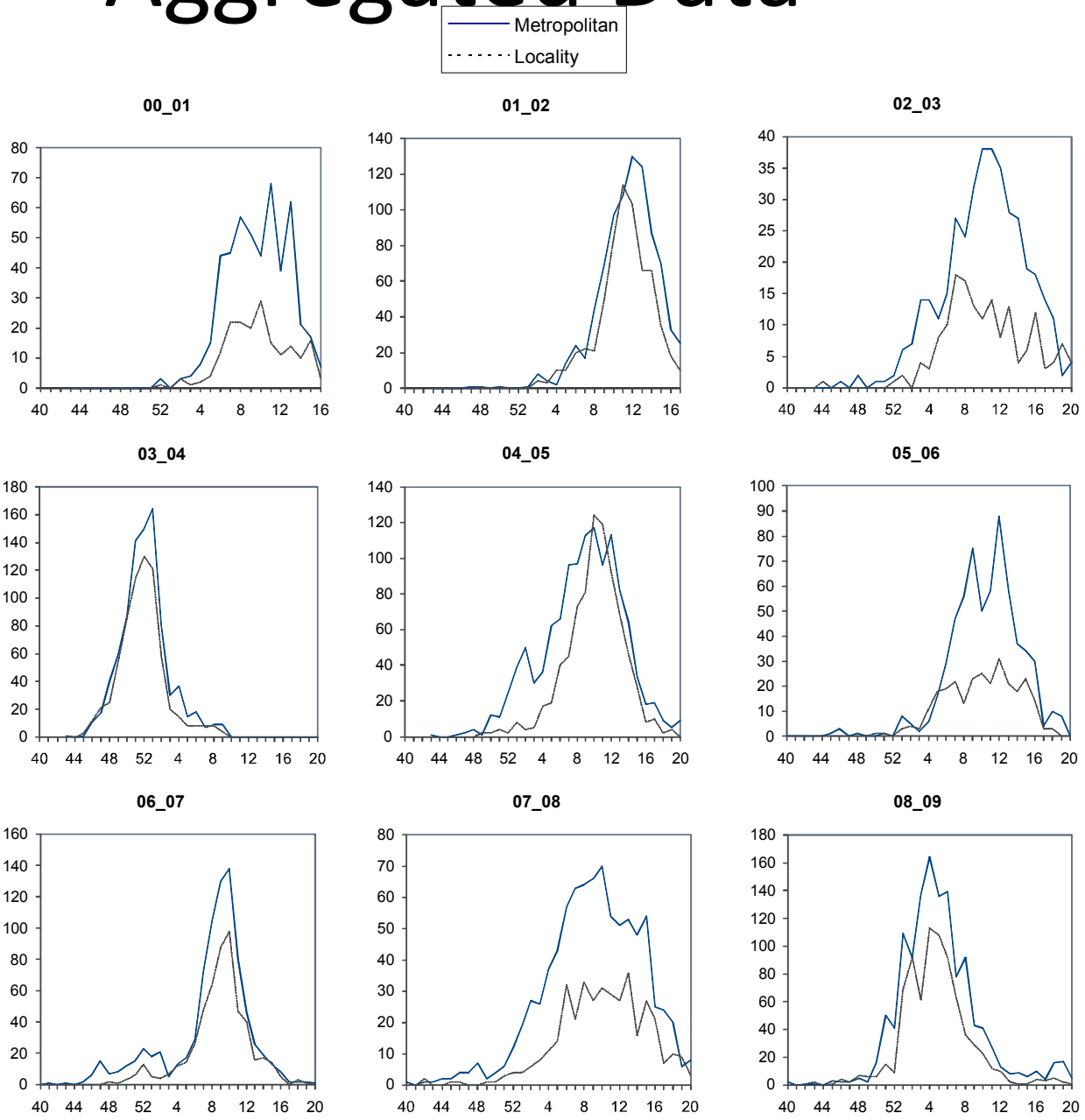
- Regional information or other data sources may be used to detect the outbreak earlier
- Time lag between regions?
  - North to south
  - East to west
  - Other

Application to Swedish Influenza

- The Swedish influenza in general starts one week earlier in the metropolitan areas than in the rest of Sweden.
- Thus a sufficient reduction may be used to detect the influenza earlier



Aggregated Data



Surveillance of Regional Data

- Data reduction incorporating time lag used
- Outbreaks detected at same time or earlier than using the whole country

Thank you for listening!

References

- Andersson, E., Kuhlmann-Berenzon, S., Linde, A. and Frisén, M., Rubinova, S., Schiöler L. (2008) Predictions by early indicators of the time and height of yearly influenza outbreaks in Sweden. *Scandinavian Journal of Public Health*, **36**, 475-482.
- Frisén, M., Andersson, E. and Schiöler, L. (2009) Robust outbreak surveillance of epidemics in Sweden. *Statistics in Medicine,* **28**, 476-493.
- Frisén, M., Andersson, E. and Schiöler, L. (2010) Evaluation of Multivariate Surveillance. *Journal of Applied Statistics*, **37**, 2089-2100.
- Frisén, M., E. Andersson and Schiöler, L. (2011). "Sufficient reduction in multivariate surveillance." Communications in Statistics -Theory and Methods **40(10): 1821-1838.**
- Schiöler, L. (2011). "Characterization of influenza outbreaks in Sweden." Scandinavian Journal of Public Health **39: 427-436**
- Schiöler, L. and M. Frisén (2012). "Multivariate outbreak detection." Journal of Applied Statistics **39(2): 223-242.**
- Frisén, M. (2014). "Spatial outbreak detection based on inference principles for multivariate surveillance." IIE Transactions **46(8): 759-769.**
- My thesis (no paywall): https://gupea.ub.gu.se/handle/2077/23951

Analytic Methodologies for Disease Surveillance Using Multiple Sources of Evidence: Overview and Bayes Net Implementation

for November 6, 2014 webinar of the International Society for Disease Surveillance

Howard Burkom
Johns Hopkins Applied Physics Laboratory

APL
JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

Multivariate Analytic Fusion of Evidence: Outline

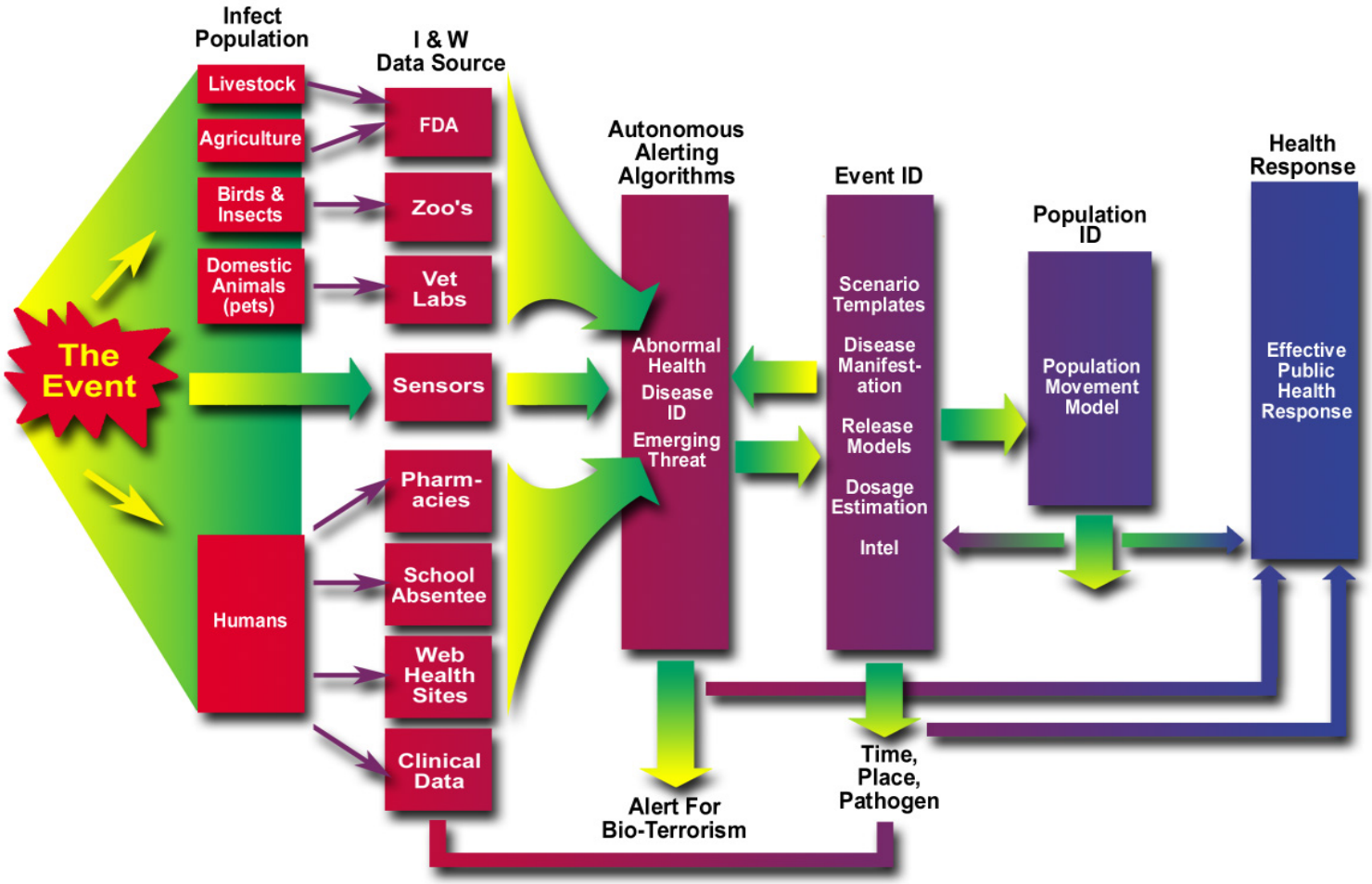
Part 1: General Remarks on Analytic Fusion of Evidence

- Concept
- What is taking so long? Obstacles
- Practical role of multivariate analytic methods
- Seeking evidenced-based validation without sufficient evidence
  - Role of simulation, historical signals, predictive models
- Many academic approaches in recent years—not enough practical collaboration

Part 2: Bayesian Networks Applied to Multivariate Syndromic Surveillance for the U.S. Department of Defense

Analytic Fusion of Evidence: Concept
Many Frameworks, Shells, Interfaces since 2000

From 1999 Program Presentation:



What is taking so long? Obstacles

- Public health monitors with the system needs do not have funding, time, manpower to design systems
  - Widespread lack of sufficient human analysts/investigators
  - Often responding to latest perceived crisis
- Wide variation in situational awareness needs
- Limited, temporary funding
- Diverse professional cultures can hinder collaboration as much as national, ethnic cultures
- Need to discern and apply appropriate use of technology: it can't do everything
  - Just because we *can* doesn't mean we *should*
- Data Availability: Need to protect patient privacy, proprietary rights, and individual and collective intellectual property

What is the Practical Role for Multivariate Analytic Methods?

Which statement most closely agrees with your perception?

1. Within the foreseeable future (say 20 years), machine learning will take over all complex decision processes involving multiple sources of evidence.
2. Certain aspects of decision making and weighting of evidence can be left to automation. The rest must remain up to the human decision maker.
3. Automated decision tools could play a useful advisory role in some situations.
4. Analytic tools can clarify data through statistical methods and data modeling but are not appropriate for decision making.
5. What is needed is not analytics, but just faster, user-friendly, and streamlined data visualization methods.

Numerous Analytic Approaches in Recent Years

Multiple Approaches, gradually converging:

Multivariate Time Series Models

- Lau E.H.Y, Cowling B.J., Ho L-M, Leung G.M., Optimizing Use of Multistream Influenza Sentinel Surveillance Data, Emerg Infect Dis. Jul 2008; 14(7): 1154–1157. doi: 10.3201/eid1407.080060

Multivariate methods based on agent-based models

- X. Jiang, G. F. Cooper, A real-time temporal Bayesian architecture for event surveillance and its application to patient-specific multiple disease outbreak detection, Data Min. Knowl. Discov. 20 (3) (2010) 328-360.

Agent-based models: computational improvements

- Skvortsov A, Ristic B, Monitoring and prediction of an epidemic outbreak using syndromic observations, Math. Biosci. (2012),

Spatiotemporal application of multivariate branching process model:

- Paul M, Held L, and Toschke AM, Multivariate modelling of infectious disease surveillance data, Statistics in Medicine, Volume 27, Issue 29, pages 6250–6267, 20 December 2008

Bayesian shared component model framework, extending SCPO:

- Corberán-Vallet A, Prospective surveillance of multivariate spatial disease data, Stat Methods Med Res. 2012 October ; 21(5): 457–477.

Nature of Required Collaboration

1. Multiple evidence sources should be used to clarify surveillance picture, not obscure it
   - Utility, effective role of new data sources derived from social media?
2. Requirements:
   - Effective visualization tools are essential
   - Transparency: epidemiologist users will not accept black-box output for decision-making
     - Analytic multivariate tools must be well explained, produce logical outputs in canonical scenarios
   - Manage data dropouts, other quality problems
   - Appropriate weighting of clinical, syndromic, environmental evidence

Part 2: Bayesian Networks Applied to Multivariate Syndromic Surveillance for the U.S. Department of Defense

Howard Burkom, Yevgeniy Elbert,

Liane Ramac-Thomas, Christopher Cuellar

Johns Hopkins Applied Physics Laboratory

Acknowledgements:

LCDR Rhonda Lizewski, Dr Julie Pavlin, Armed Forces Health Surveillance Center

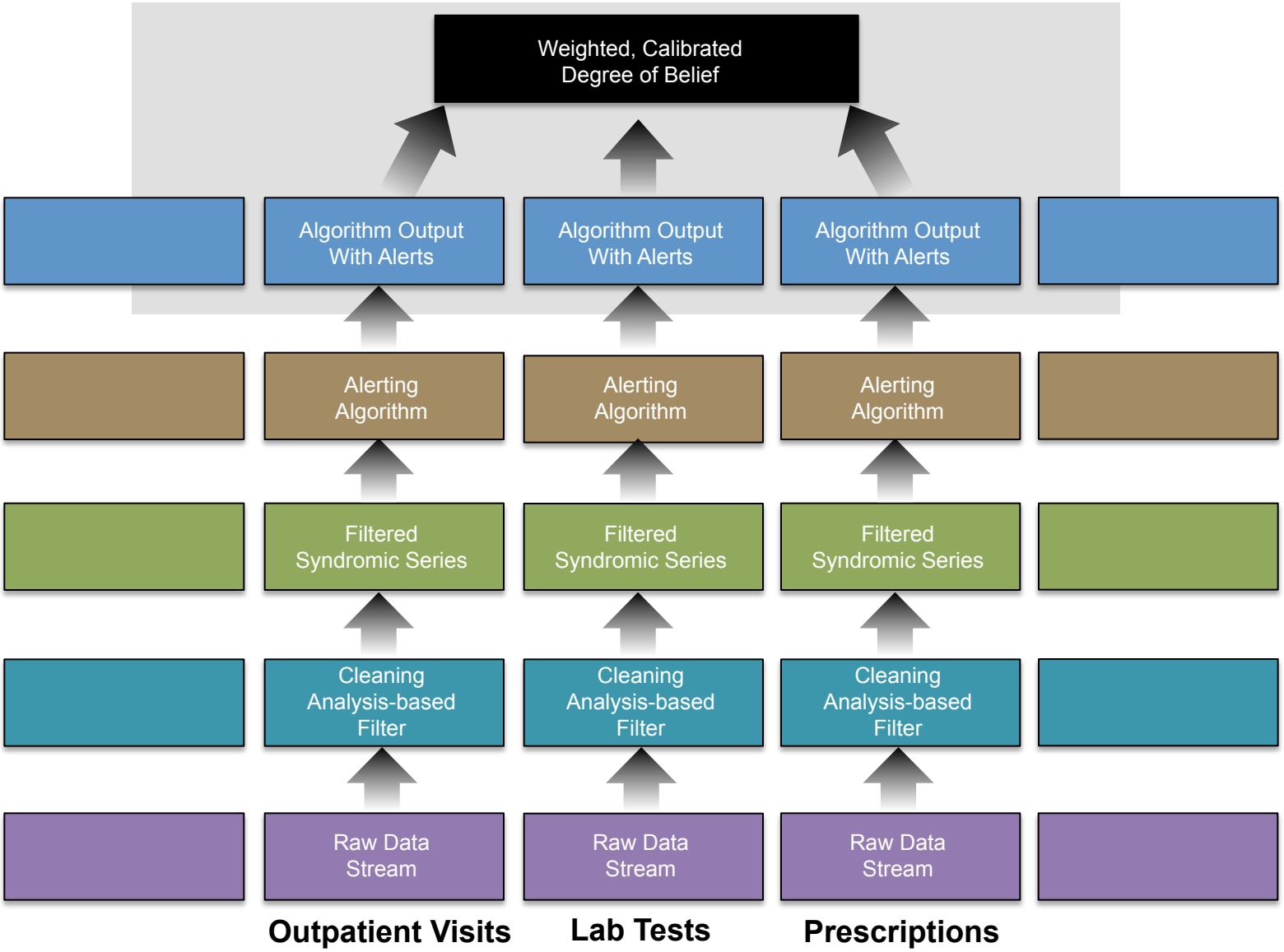
Joe Lombardo, JHU/APL

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

Bayes Networks for DoD Surveillance: Outline

- Scope, Evidence Sources of DoD Surveillance
- Concept: Bayesian Networks for Analytic Fusion
- Implementation Summary
- Recent Results
- Discussion: validation and practical usage

From Separate Algorithms to Integrated Decision Support for DoD Surveillance



Development Data Environment

- Historical Dataset of 3.75 years, all US military treatment facilities
  - Data Sources
    - Outpatient records, including ICD-9, chief complaints, demographic, severity-related fields
    - Chemistry and Microbiology Laboratory test orders & results
    - Filled Prescriptions
  - Data dates: 1360 continuous days, 10Jan2007 – 29Sep2010
- Data from 502 individual treatment facilities
  - Including 289 hospitals and large clinics with all data sources
- Truth data: reported outbreaks in three clinical categories
  - Influenza-like illness (ILI)
  - Gastrointestinal illness (GI)
  - Febrile illness (Fever)

References:

[MSMR] Medical Surveillance Monthly Report, April 2012, Volume 19, Number 4, http://www.afhsc.mil/viewMSMR?file=2012/v19_n04.pdf

[TRICARE] Defense Health Cost Assessment and Program Evaluation (DHCAPE), in the Office of the Assistant Secretary of Defense (Health Affairs) (OASD/HA) (2012) Evaluation of the TRICARE Program: Fiscal Year 2012 Report to Congress. http://www.tricare.mil/hpae/_docs/TRICARE2012_02_28v5.pdf.

[DHSS ESSENCE] http://www.health.mil/Military-Health-Topics/Technology/Decision-Support/Electronic-Surveillance-System-for-the-Early-Notification-of-Community-based-Epidemics

Approach to Analytic Fusion of Evidence

Instead of:

- how to model data effects of public health threats,
- how evidence sources are correlated,
- which data signals correspond to authentic events and false alarms,
- how to analytically combine data from different, weighted data sources,

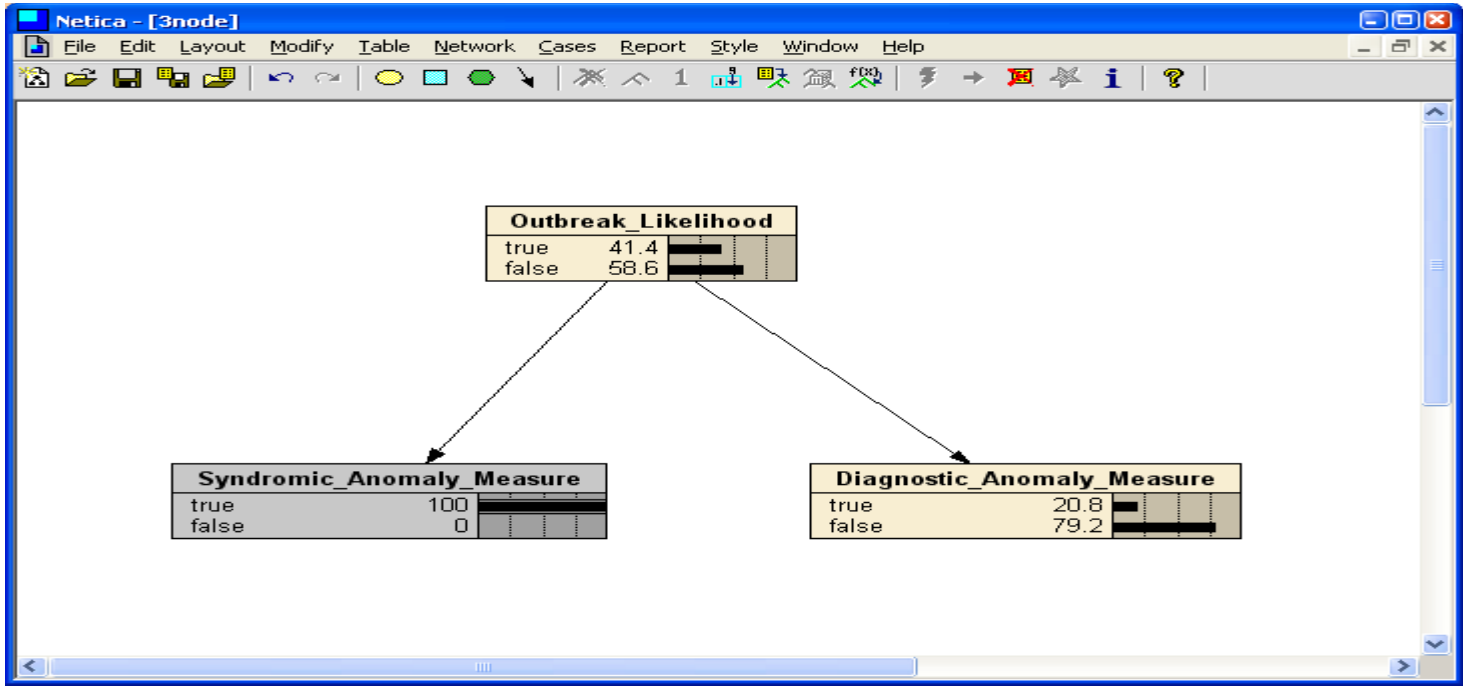
We ask:

- how would an experienced health monitor make investigation decisions given the luxury of examining all data sources every day?

Concept: Population-based Bayes Networks

- Method of combining information from the monitored population
  - Algorithm results from multiple data streams of varying relevance (not raw data)
  - More than a rule set: an analytic umbrella that can also include report-based results, incomplete data updates, other multivariate methods
- Not Bayesian statistics in the sense of hierarchical modeling, fixed/random effects (could incorporate)
- Not an agent-based Bayesian model representing every individual as a separate node with properties

Fusing Syndromic and Definitive Evidence



For technical approach description, see:

Burkom H, Ramac-Thomas L, Babin S, Holtry R, Mnatsakanyan Z, Yund C, 2011: An integrated approach for fusion of environmental and human health data for disease surveillance. Statistics in Medicine, 30(5):470-479

Quantitative Effect of Fused Evidence

A	Outbreak occurring	
B1	Syndromic outbreak evidence	
B2	Diagnostic outbreak evidence	

Prior Outbreak Degree of Belief

Pr(A)	Pr(~A)
0.01	0.99

Conditional Probability Tables

Conditional Probability Tables		
	cond Pr(B1)	cond Pr(~B1)
A	0.7	0.3
~A	0.01	0.99

	cond Pr(B2)	cond Pr(~B2)
A	0.7	0.3
~A	0.001	0.999

Degree of Belief Given Two Evidence Sources

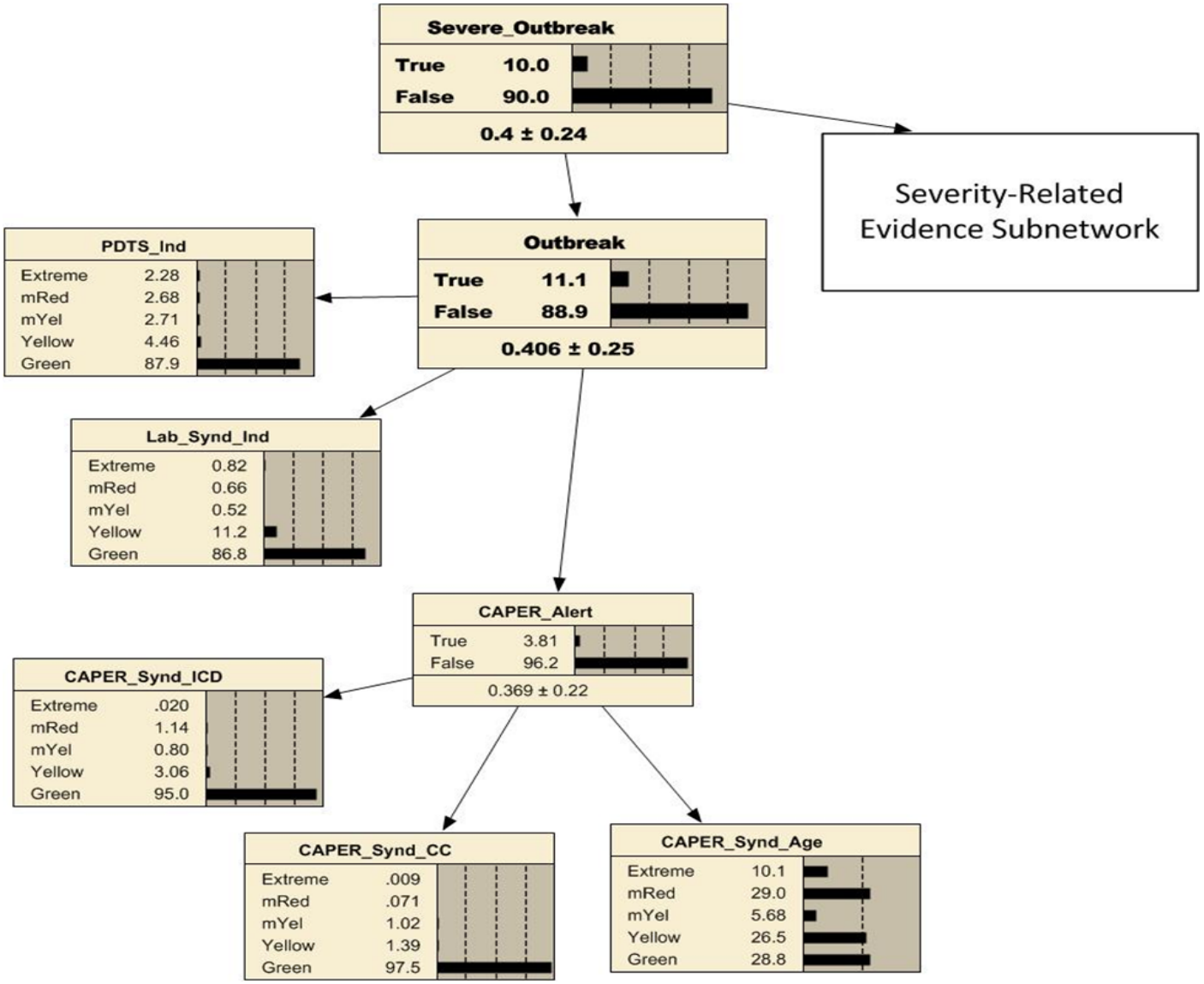
Pr(A,B1,B2)	0.00490
Pr(~A,B1,B2)	0.00001
Pr(B1,B2)	0.00491
Pr(A|B1,B2)	0.99798

Degree of Belief Given One Evidence Source

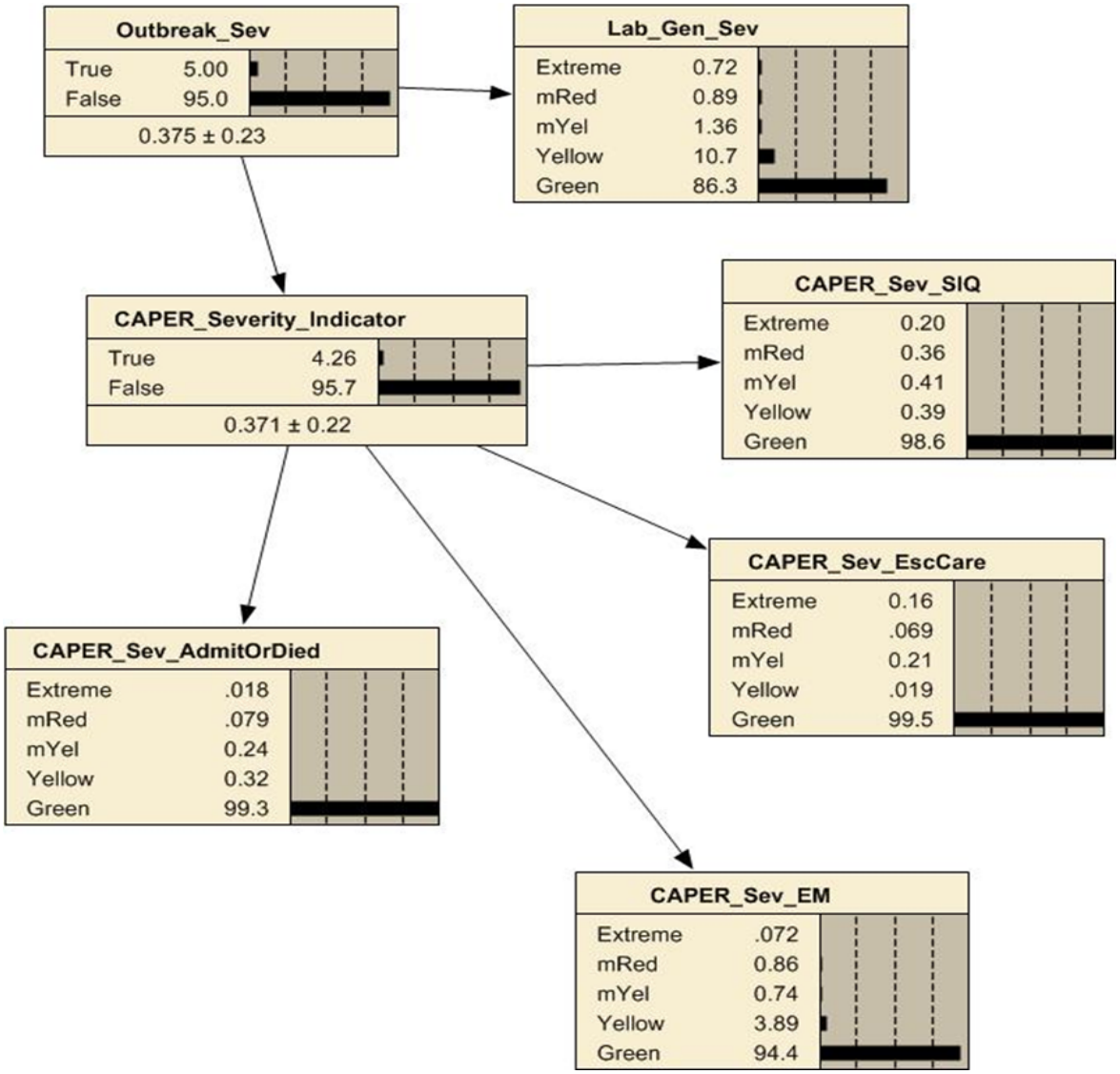
Pr(B1,A)	Pr(B1,~A)	Pr(B1)	Pr(A|B1)
0.00700	0.00990	0.01690	0.41420

Pr(B2,A)	Pr(B2,~A)	Pr(B2)	Pr(A|B2)
0.00700	0.00099	0.00799	0.87610

Generic Nodal Structure for Fusion



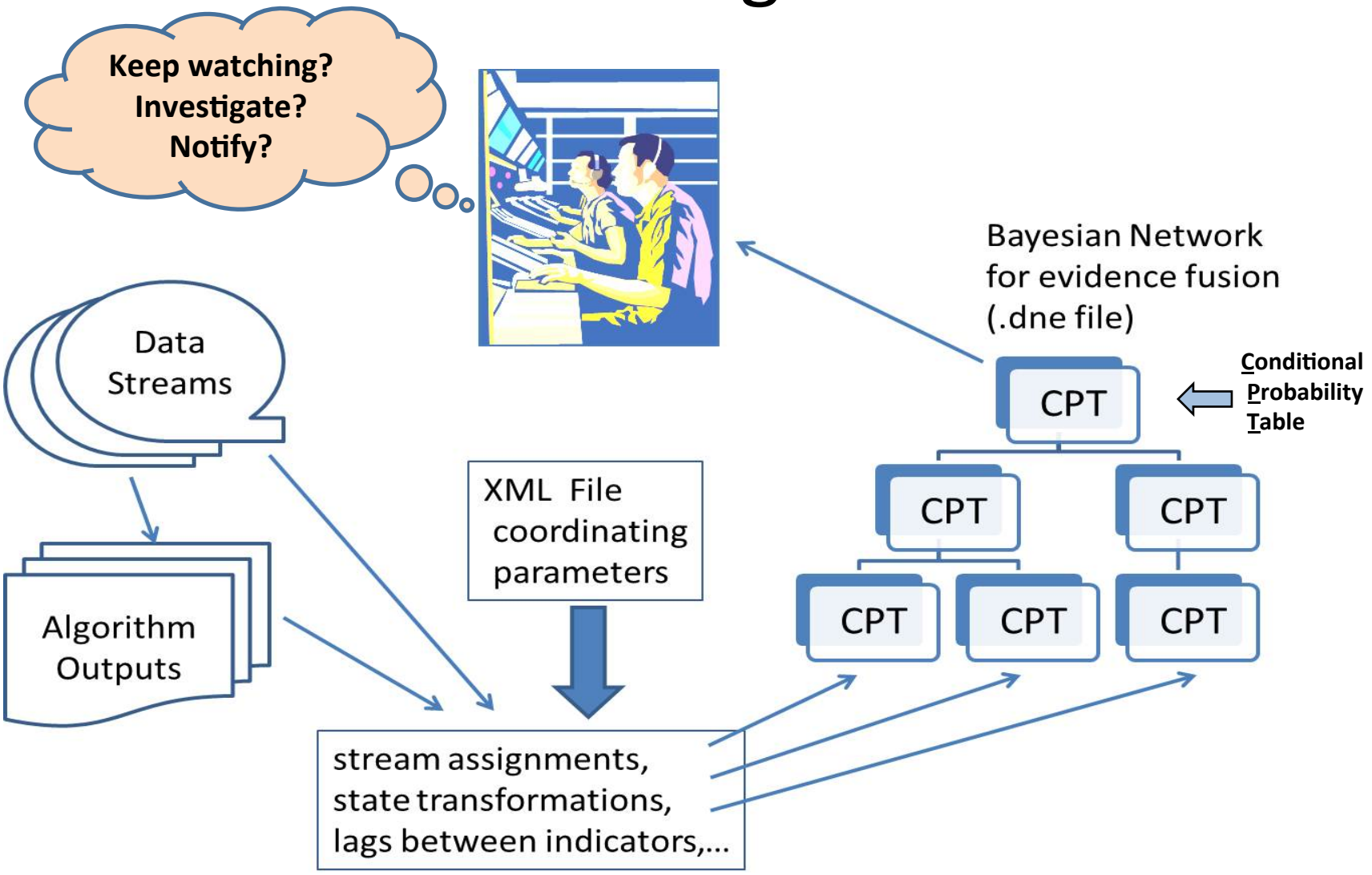
Bayesian Fusion: Severity Sub-Network



Overview of Machine-Learning Methods

- Network structures: mainly heuristic, based on guidance from medical epidemiologist
- Conditional probability tables for each node: optimized to yield desired degree of concern for canonical input state combinations
  - dependent on input from epidemiologist domain experts
- Calibration: multivariate search to produce combination of thresholds for indicator algorithm states and for network decision nodes
  - detect all known events with highest decision node odds ratios
- Validation:
  - checked performance on 30 known health events,
  - 10-fold cross-validation based on undocumented, data-derived events

Schematic: ESSENCE Fusion Alerting



Alerting Summary for Influenza-like Illness Fusion Syndrome: 289 Facilities over 3.75 Years

Cumulative ILI Alerting: 10Jan2007 - 29Sep2010 (3.75 yrs)			289 Facilities		50 Largest Facilities	
	ILI Fusion Network and Main Indicator Streams	mean alerts/ day: all facilities	sum of all alerts	mean days betw. alerts per facility	sum of all alerts	mean days betw. alerts per facility
General Indicator Data Streams	CAPER ICD-9 Syndrome	6.9	9018	41.7	2997	21.7
	CAPER Chief Complaint Syndrome	7.8	10145	37.0	2181	29.8
	CAPER Age<18	6.1	7914	47.5	1688	38.5
	Syndromic Lab Test Order Group	7.0	9095	41.3	1902	34.2
	Syndromic Prescription Group	7.1	9209	40.8	1966	33.1
Severity-related Indicators	CAPER: Admitted/Died	0.6	743	505.7	239	272.0
	CAPER: Complex E/M Codes	6.1	7965	47.2	2099	31.0
	CAPER: Escalated Care	9.7	12587	29.8	2971	21.9
	CAPER: Sick-in-quarters	9.9	12911	29.1	3064	21.2
	Lab Tests/General Severity	2.3	2927	128.4	692	93.9
	All Severity-Related Indicators	28.6	37133	10.1	9065	7.2
Influenza Indicator Data Streams	Influenza Antivirals	7.1	9247	40.6	1440	45.1
	Positive Influenza Lab Tests	0.6	835	449.9	289	224.9
All Streams	All Data Indicators	71.2	92596	4.1	21528	3.0
Fusion Network Decision Nodes	**General Outbreak Fusion**	**2.5**	**3239**	**116.0**	**937**	**69.4**
	Severity-Related Outbreak Fusion	0.8	1014	370.5	322	201.9
	Influenza Outbreak Fusion	0.3	432	869.7	158	411.4

Odds Ratio Analysis of Indicator and Fusion Outputs for Reported ILI Events

Odds Ratios for Reported ILI Events		Indicator Streams											Fusion Decision Nodes		
Facility or Region	Dates	CAPER ICD-9 Syndrome	CAPER Chief Complaint Syndrome	Syndromic Lab Test Order Group	CAPER Age<18	CAPER: Admitted/Died	CAPER: Complex E/M Codes	CAPER: Escalated Care	CAPER: Sick-in-quarters	Lab Tests/General Severity	Syndromic Prescription Group	Influenza Antivirals	Positive Influenza Lab Tests	Severity-Related Outbreak Fusion	General Outbreak Fusion	Influenza Outbreak Fusion
	2/1/2009-2/21/2009	3.9	0.0	0.0	0.0	N/A	5.0	1.6	1.4	0.0	0.0	2.2	0.0	N/A	**19.1**	N/A
	9/28/2009-12/31/2009	**28.8**	9.4	0.0	1.9	N/A	9.3	6.0	5.7	0.0	13.5	N/A	N/A	N/A	20.2	N/A
	7/3/2009-7/15/2009	21.0	23.6	6.1	N/A	56.1	19.2	23.4	23.4	N/A	0.0	0.0	N/A	0.0	**89.8**	N/A
	9/28/2009-12/1/2009	0.0	**7.5**	0.0	2.8	N/A	1.8	5.1	5.3	N/A	2.1	7.3	N/A	0.0	2.0	N/A
	1/27/2008-3/24/2008	0.0	2.9	**7.1**	0.0	1.2	0.0	6.5	6.1	0.0	0.0	2.0	4.2	11.4	4.8	22.8
	1/6/2008-3/11/2008	0.3	0.8	**3.4**	0.0	N/A	0.8	2.0	1.9	0.0	0.6	8.7	11.4	2.2	1.5	11.4
	7/6/2009-7/20/2009	16.0	19.2	20.7	0.0	N/A	16.4	7.1	7.5	0.0	16.6	**33.9**	0.0	144.1	30.0	21.3
Average Odds Ratios		**10.0**	**9.0**	**5.3**	**0.8**	**28.7**	**7.5**	**7.4**	**7.3**	**0.0**	**4.7**	**9.0**	**3.9**	**31.5**	**23.9**	**18.5**
Median Odds Ratios		**3.9**	**7.5**	**3.4**	**0.0**	**28.7**	**5.0**	**6.0**	**5.7**	**0.0**	**0.6**	**4.7**	**2.1**	**2.2**	**19.1**	**21.3**
Events detected (/7)		**5**	**6**	**4**	**2**	**2**	**6**	**7**	**7**	**0**	**4**	**5**	**2**	**3**	**7**	**3**

Alerting

and I vents

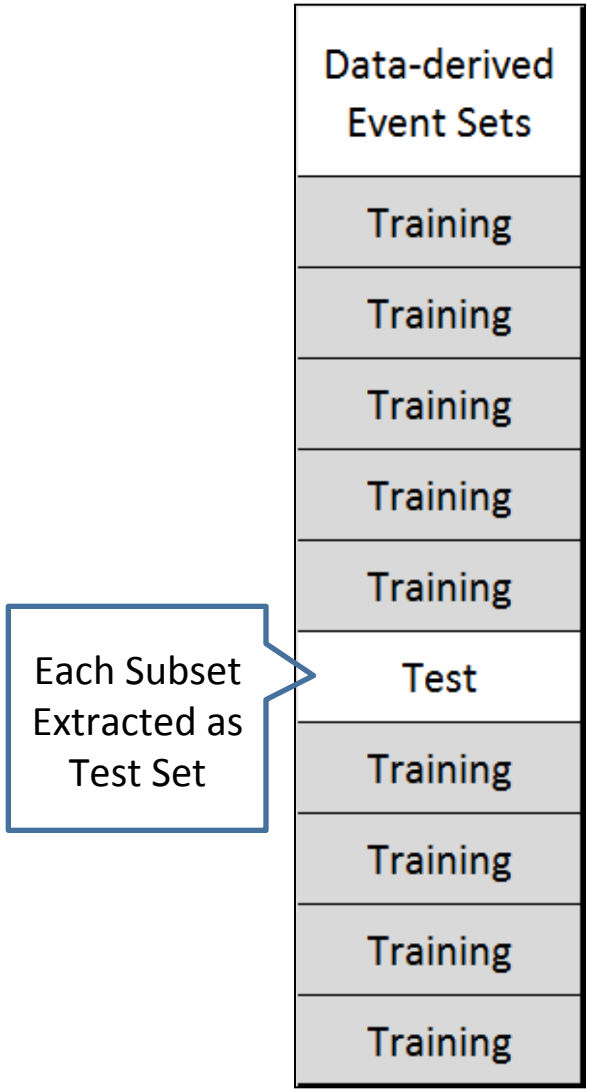
Nr.	Syndrome	Facility	Date of First Fusion Alert	Date of First ICD Algorithm Alert	Alerting Lag (days) +: ICD first, -: fusion first	ICD indicator alert first	Fusion Network alert first	Fusion/ICD Same Day Alerting
1	ILI		1/4/2008	No ICD alert	(fusion only)	0	1	0
2	ILI		2/3/2008	2/18/2008	-15	0	1	0
3	ILI		7/6/2009	7/7/2009	-1	0	1	0
4	ILI		2/15/2009	2/15/2009	0	0	0	1
5	ILI		7/8/2009	7/5/2009	3	1	0	0
6	ILI		10/8/2009	9/28/2009	10	1	0	0
7	ILI		2/1/2008	1/6/2008	26	1	0	0
			ILI Events			**3**	**3**	**1**
8	FEVER		5/1/2007	5/3/2007	-2	0	1	0
9	FEVER		6/21/2007	6/21/2007	0	0	0	1
10	FEVER		7/6/2009	7/6/2009	0	0	0	1
11	FEVER		3/17/2009	3/16/2009	1	1	0	0
12	FEVER		1/31/2008	1/29/2008	2	1	0	0
13	FEVER		1/14/2008	1/3/2008	11	1	0	0
14	FEVER		10/31/2009	10/13/2009	18	1	0	0
15	FEVER		2/1/2008	1/9/2008	23	1	0	0
			Fever Events			**5**	**1**	**2**
16	GI		1/19/2010	1/26/2010	-7	0	1	0
17	GI		1/20/2010	1/25/2010	-5	0	1	0
18	GI		1/9/2010	1/12/2010	-3	0	1	0
19	GI		3/10/2010	3/10/2010	0	0	0	1
20	GI		2/22/2010	2/22/2010	0	0	0	1
21	GI		3/25/2010	3/25/2010	0	0	0	1
22	GI		2/17/2010	2/17/2010	0	0	0	1
23	GI		6/12/2009	6/12/2009	0	0	0	1
24	GI		6/18/2007	6/18/2007	0	0	0	1
25	GI		1/4/2008	1/4/2008	0	0	0	1
26	GI		8/7/2009	8/7/2009	0	0	0	1
27	GI		12/26/2009	12/26/2009	0	0	0	1
28	GI		1/26/2010	1/26/2010	0	0	0	1
29	GI		10/28/2008	10/27/2008	1	1	0	0
30	GI		11/15/2007	11/13/2007	2	1	0	0
			GI Events			**2**	**3**	**10**
			All Events			**10**	**7**	**13**

The Validation Problem: "Why adopt this?"

"Why should you believe me?"

- Why believe that sensitivity to a few dozen historical events promises sensitivity to future events? How to get acceptance?
- Burden of proof is discipline-specific
  - Public health epidemiology
    - authentic outbreak effects difficult to ascertain in data
    - much more difficult in multivariate data
  - Computer science/data mining
    - Simulations widely used, highly developed, but problematic
    - Relative strength of signal, timing across datasets

Ten-Fold Cross-validation using Derived Events from Authentic Data



- Unreported "events" derived from 3.75 years of data from 289 facilities
  - Corroboration in syndromic, clinical data
  - 101 events for Fever, 73 for GI, 128 for ILI
- Events partitioned into 10 subsets of facilities
- Ten sensitivity tests: one subset removed
  - Other 9 subsets pooled for training
  - Optimal combination of thresholds found
  - Network with optimal thresholds applied to test subset
- RESULTS: 95-99% of extracted "events" detected for each syndrome-specific Bayesian Network

Summary

- Cross-validation method supports use of Bayesian Network fusion approach to combine multiple data sources for prospective alerting
- Learning to combine historical data analysis with knowledge of experienced medical epis
- Next step: application to true real-time alerting
- Essential for realization of fusion capability
  - Training
  - Visualization
  - Integration with other surveillance tools

Toward Portable, Translatable Fusion of Surveillance Evidence Fusion: Technical Challenges

- Difficulty of insufficient truth data, especially outbreak effects in multiple indicators:
  - Relative timing, relative degree of effects in different data??
- At what level and consistency is a purely syndromic signal worthy of investigation?
- All data sources contaminated in some way
  - Challenge to extract meaningful, representative indicators
  - Need right balance between granular analysis and robust data behavior
- Expert elicitation problem
  - Judgment of medical epidemiologists
  - Extrapolation to all state combinations

Thank you for your attention!

Howard.Burkom@jhuapl.edu

Summary remarks

Marianne Frisén
University of Gothenburg

Detection of outbreak of health threat.

Enhancement by use of multiple sources.

Relations between different variables are examined and utilized:

Combination of evidence from sources of different kinds (health and environment)

versus

Optimal combination of data at a time lag between incidence in different geographical areas

Inferential approach

Bayes –Summarize different kinds of information into a measure of the probability that there is a threat.

versus

Frequentistic - Good properties in the long run

Health threat

Influenza

Increasing

Data summarized over time

Waterborne disease

Worse than baseline

Data evaluated separately at each time

Combine information

- Outbreak alarm

- Epidemiologic experience
- Visualization

Upcoming ISDS Events

- Webinar: November 12, 2014 – Planning for the ICD-10 Transition
- November 7, 2014 – Meaningful Use Communtiy Call
- Webinar: November 20, 2014 - Animal Surveillance in the US

Thank you for attending!

ISDS
INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE