Meeting Challenges in the Use of Nontraditional Data Sources for Public Health Monitoring

Yevgeniy Elbert
Walter Reed Army Institute for Research
Department of Preventive Medicine

Howard Burkom
National Security Technology Department,
John Hopkins University Applied Physics Laboratory

2nd Annual Public Health Information Network Stakeholders' Conference

Atlanta, GA May 26, 2004

Outline of Presentation

- Developing standards for evaluating a data source for surveillance
- Example: assessing effect of late reporting
- Measurement of lateness effect
- Adjustment for reporting lag
  - Use of historical reporting patterns
  - Inference from counts of reporting providers

Developing standards for evaluating a data source for surveillance

Need For Objective Assessment Standards :

- Does it provide an early or corroborative indicator of the presence of an outbreak?
- Can unrelated sources of statistical anomalies be modeled or otherwise "explained away" to reduce the nuisance factor of false alerts?
- Are up-to-date counts available on a timely basis?

Reporting: Visits from 1 Civilian Source

Civilian Resp. Synd.	Reporting Lag in Days														
Sum of Visit Counts	Lag														
Visit Date	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
07/01/03	28	36	82	0	0	0	57	35	53	137	75	0	0	51	32
07/02/03	7	98	0	0	0	64	37	47	96	72	0	0	43	56	26
07/03/03	47	0	0	0	49	59	71	48	87	0	0	36	47	55	27
07/04/03	0	0	0	10	1	13	6	12	0	0	30	13	2	11	15
07/05/03	0	0	12	13	13	11	10	0	0	42	5	12	1	40	0
07/06/03	0	2	3	1	4	7	0	0	23	11	2	3	32	0	0
07/07/03	0	95	102	67	78	0	0	122	57	43	27	58	0	0	13
07/08/03	21	79	63	64	0	0	107	69	51	56	80	0	0	19	40
07/09/03	18	66	53	0	0	90	52	48	65	81	0	0	5	39	17
07/10/03	21	50	0	0	89	84	69	44	100	0	0	21	48	16	11
07/11/03	42	0	0	94	51	76	50	91	0	0	10	60	21	13	9
07/12/03	0	0	26	4	14	14	35	0	0	20	16	8	3	10	0
07/13/03	0	8	1	4	0	20	0	0	22	17	8	8	4	0	0
07/14/03	35	81	96	72	94	0	0	26	129	76	34	22	0	0	10

No July 4 Reporting

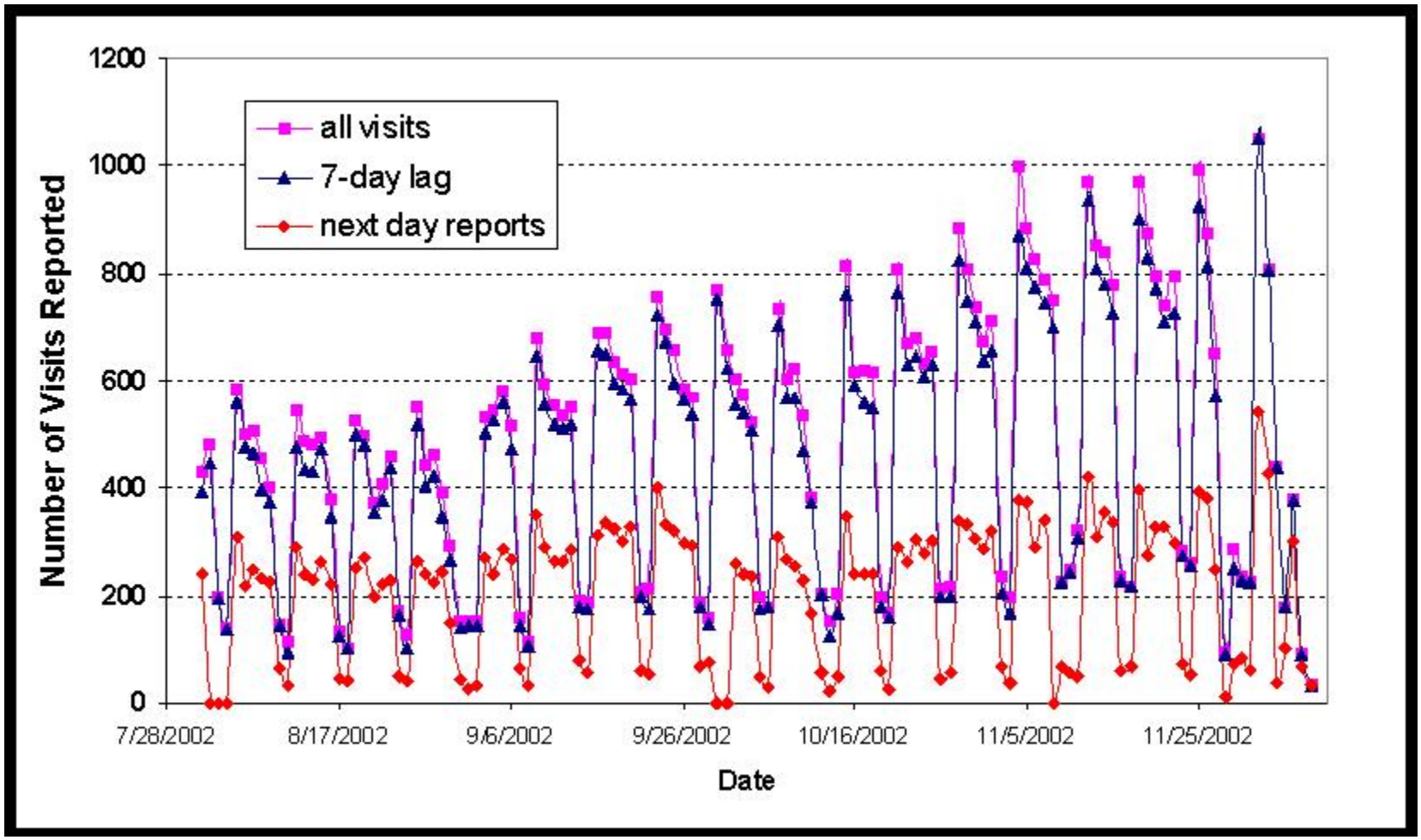
Sundays in red

Late Reporting: the Monitor's Dilemma

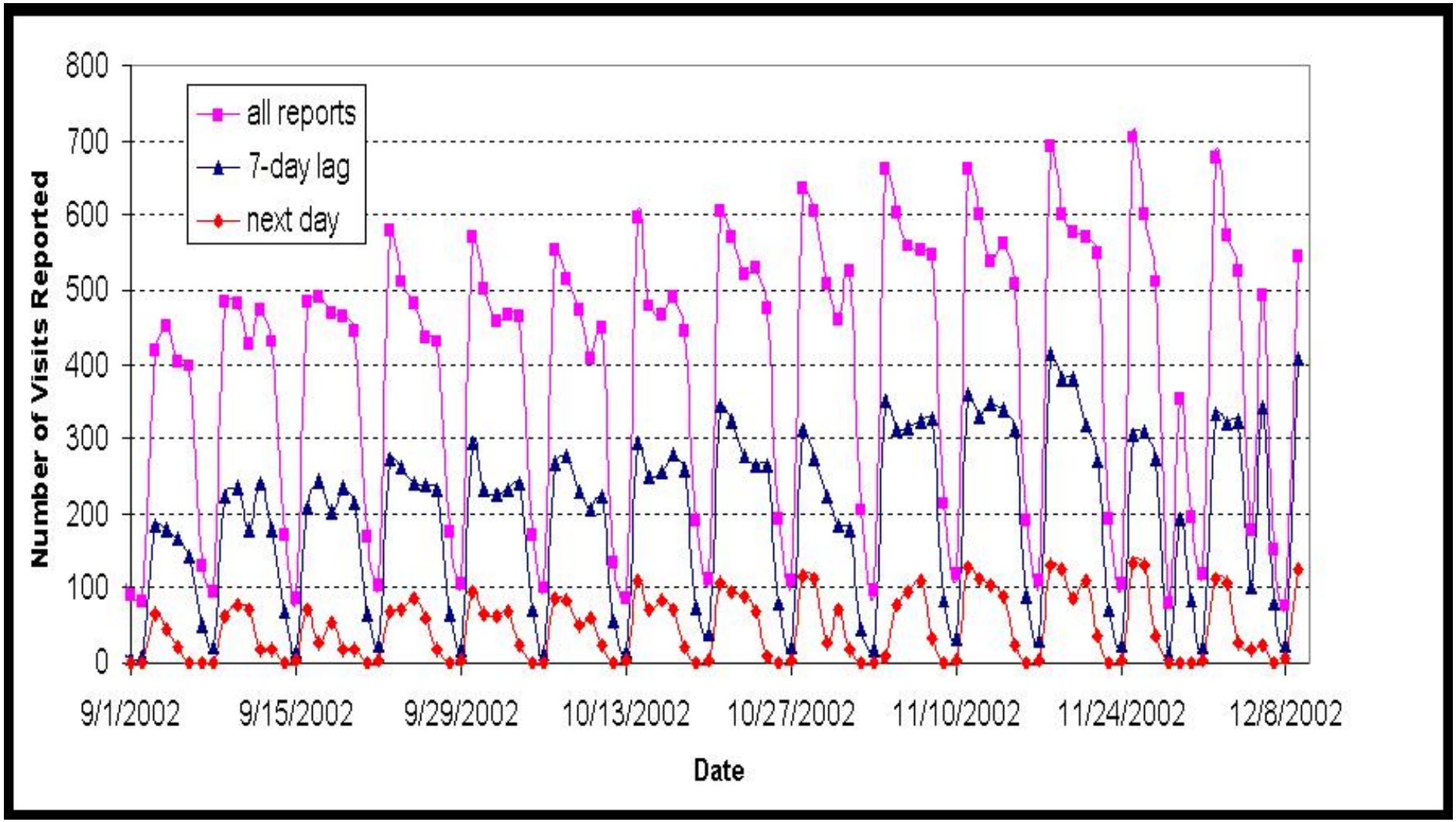
How to use late-reported data?

- Backfilling uses all of the data,
  - but recent data under-represent actual counts,
  - and recent actual counts are of greatest interest!
- Limiting use of backfill: restriction to data reported early (say within 4 days)
  - Captures time series behavior of actual counts IF reporting is consistent
  - Day-to-day experience => ignoring later reports will produce false alerts, may miss important ones
- Use inference to adjust for recent late reporting

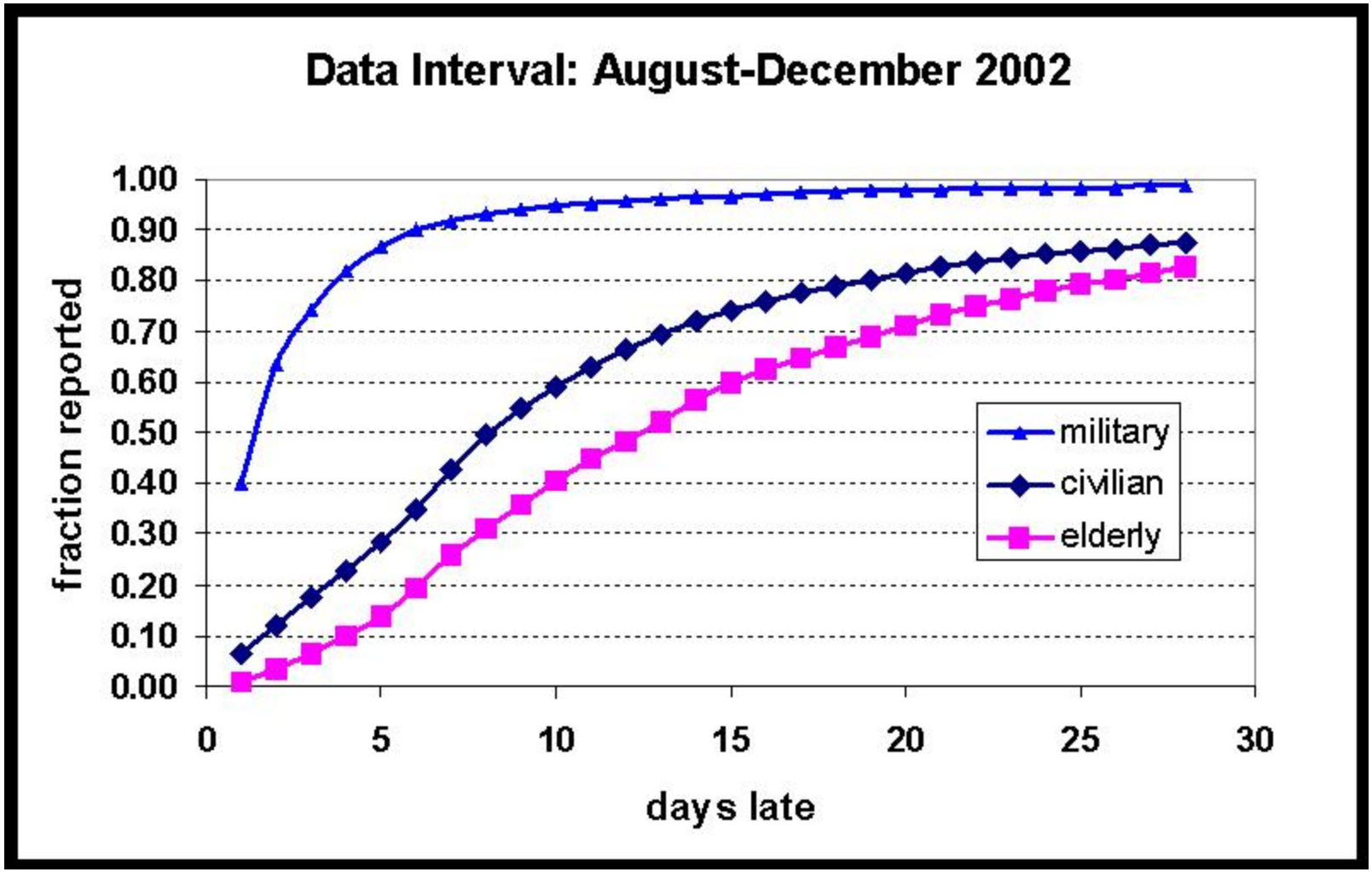
Reporting of Military Office Visits



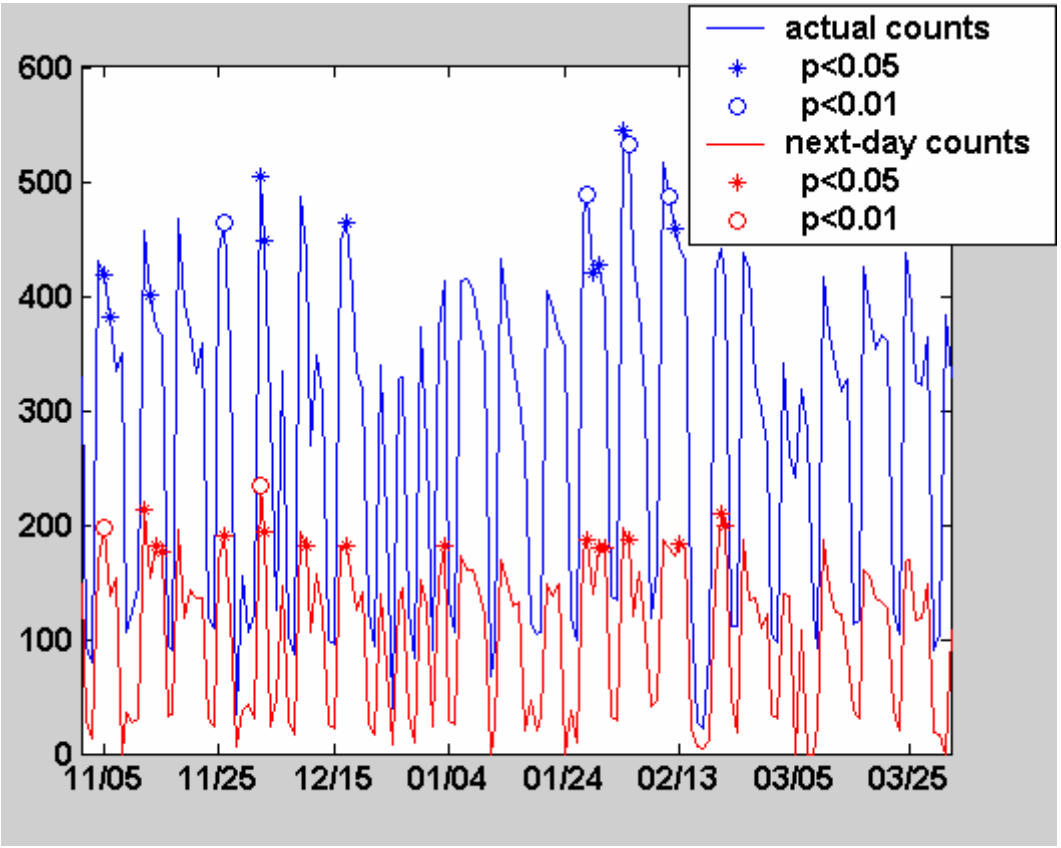
Reporting of Civilian Office Visits



Office Visit Reporting Promptness by Data Source



Measuring Surveillance Cost of Late Reporting



Counts of Algorithmic Results

Actual \ Next-Day	p< 0.01	p<0.05	no flag	
p<0.01	0	3	1	4
p<0.05	2	4	11	17
no flag	0	10	397	407
	2	17	409	428

Kappa Statistic: κ = (Po – Pe) / (1 – Pe)

Pe = proportion of observed agreement

Po = proportion of expected agreement

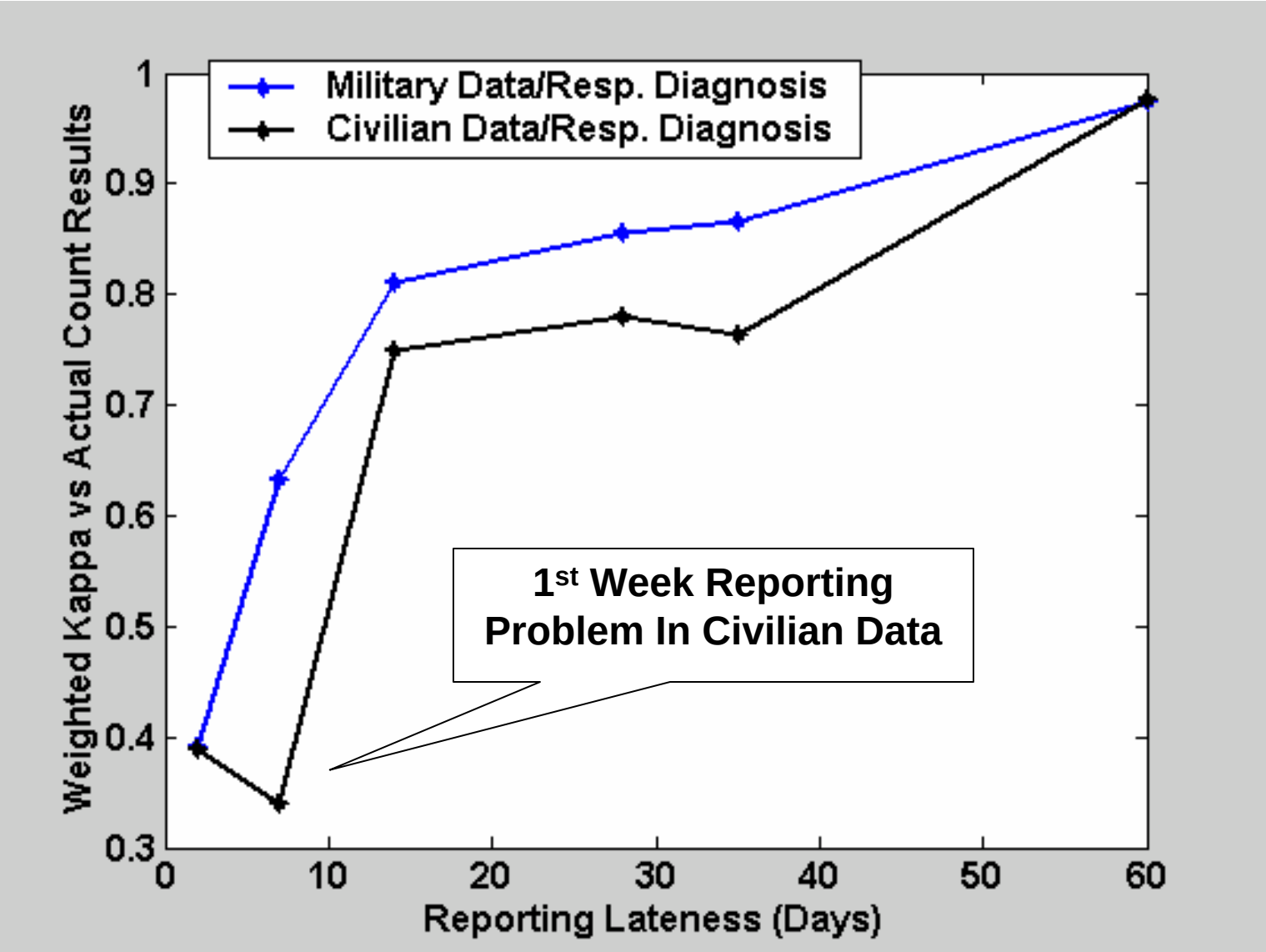
κ = 0.30

weighted κ = 0.39

(0.75 for disagreement over anomaly significance)

Data Source Lateness Comparison

Weighted Kappa Comparing Regression Results



Using Lagged Data Counts for Biosurveillance

- ESSENCE data => hypothesis that earlier stages of an outbreak may be more detectable in office visit (OV) data than in emergency department data
  - Depends on existence, duration of typical prodrome for underlying disease
  - How to exploit this for earlier alerting?
- BUT, our electronic OV data is reported variably late, depending on individual providers
- QUESTION: How can a timely source of data with a reporting lag be used for biosurveillance?

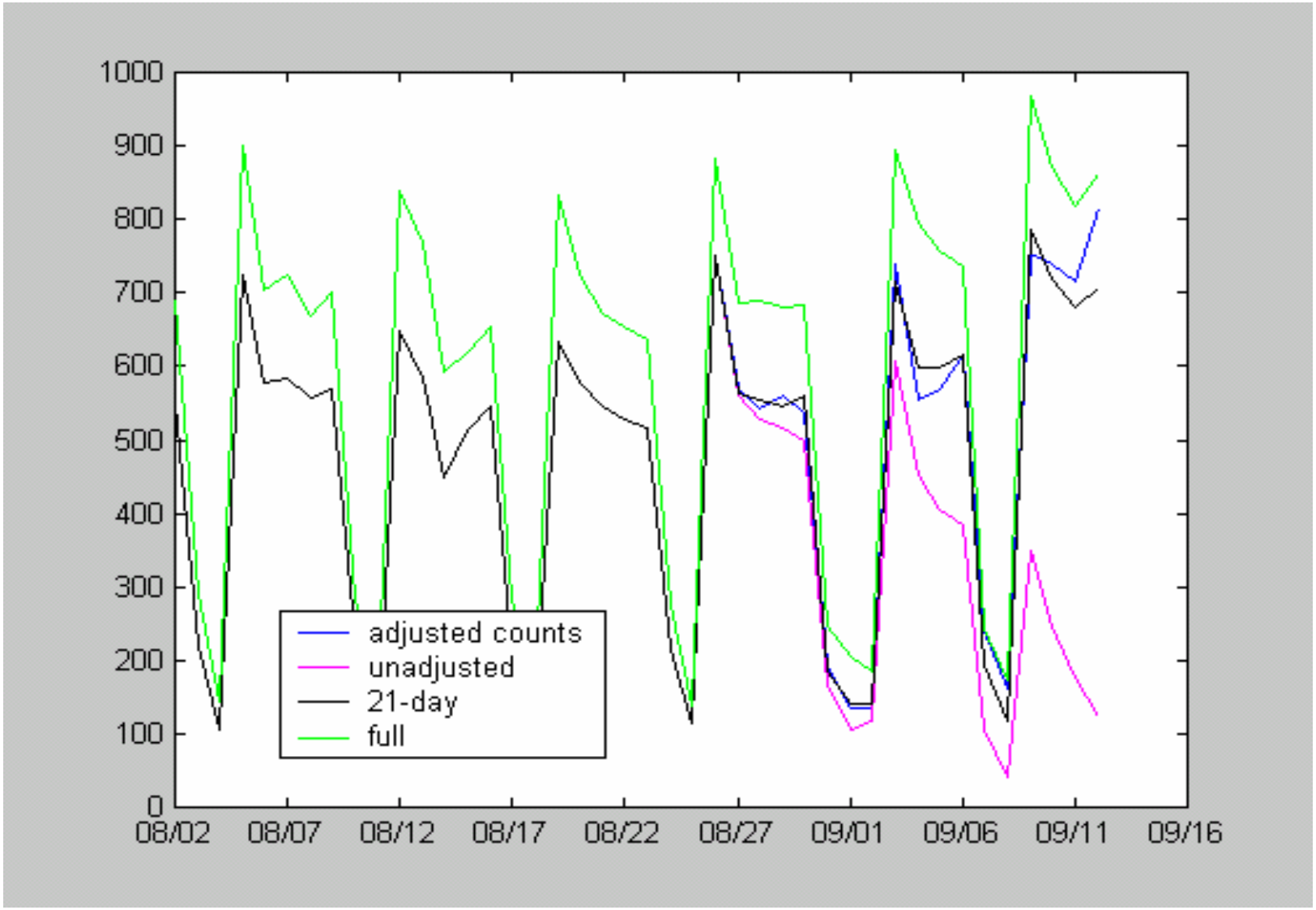
Using Lagged Data for Biosurveillance Approaches

- Two steps: estimate actual counts, apply algorithm
  - use recent promptness functions by day-of-week, other covariates
  - apply lateness factors to recent counts

  Brookmeyer R, Gail MH, *AIDS Epidemiology: A Quantitative Approach*. New York: Oxford University Press; 1994; Chapter 7
- Use historically early reporting providers as sentinels
- Combined approach: use regression on counts with date and lag as predictors to determine whether recent reported data are anomalous

  Zeger, SL, See, L-C, Diggle, PJ, "Statistical Methods for Monitoring the AIDS Epidemic", *Statistics in Medicine* 8 (1999)
- Regression including number of providers reporting each day

Reporting of Civilian Office Visits
21-day adjustment Week 1

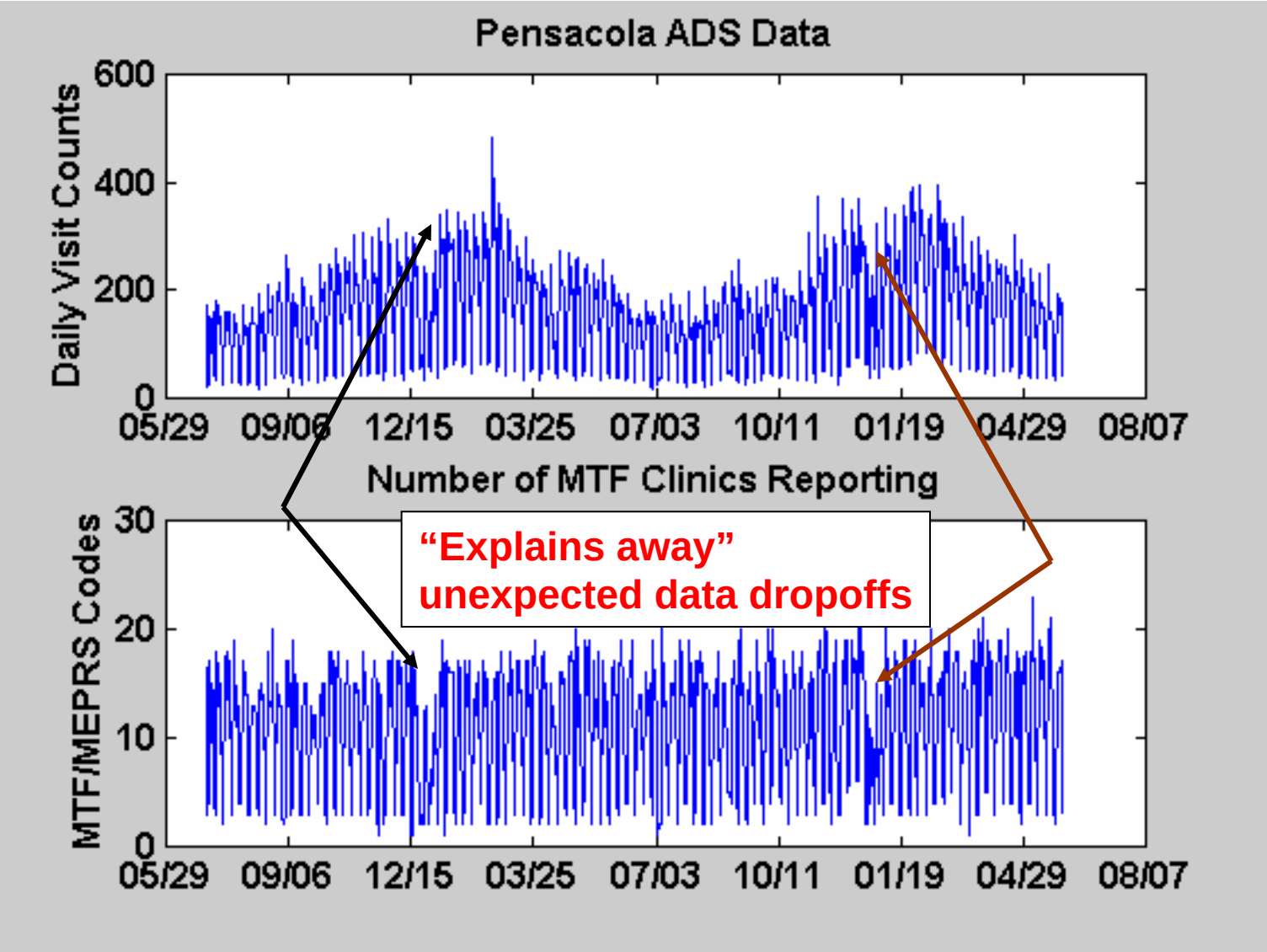


Use of Daily Number of Reporting Providers to Reduce Lag Effect

- Concept:
  - Tabulate daily counts of providers reporting in dataset: physician IDs, clinics, pharmacies
  - Include these provider counts as predictors in regression of daily visit counts
  - Account for reporting lag as well as known & unknown dropoffs by computing actual counts vs expected, given number of providers reporting
- Can apply process control algorithms to residuals
- Significantly attenuates day-of-week effect

Counts of Clinic/MTF Pairs

ADS Data

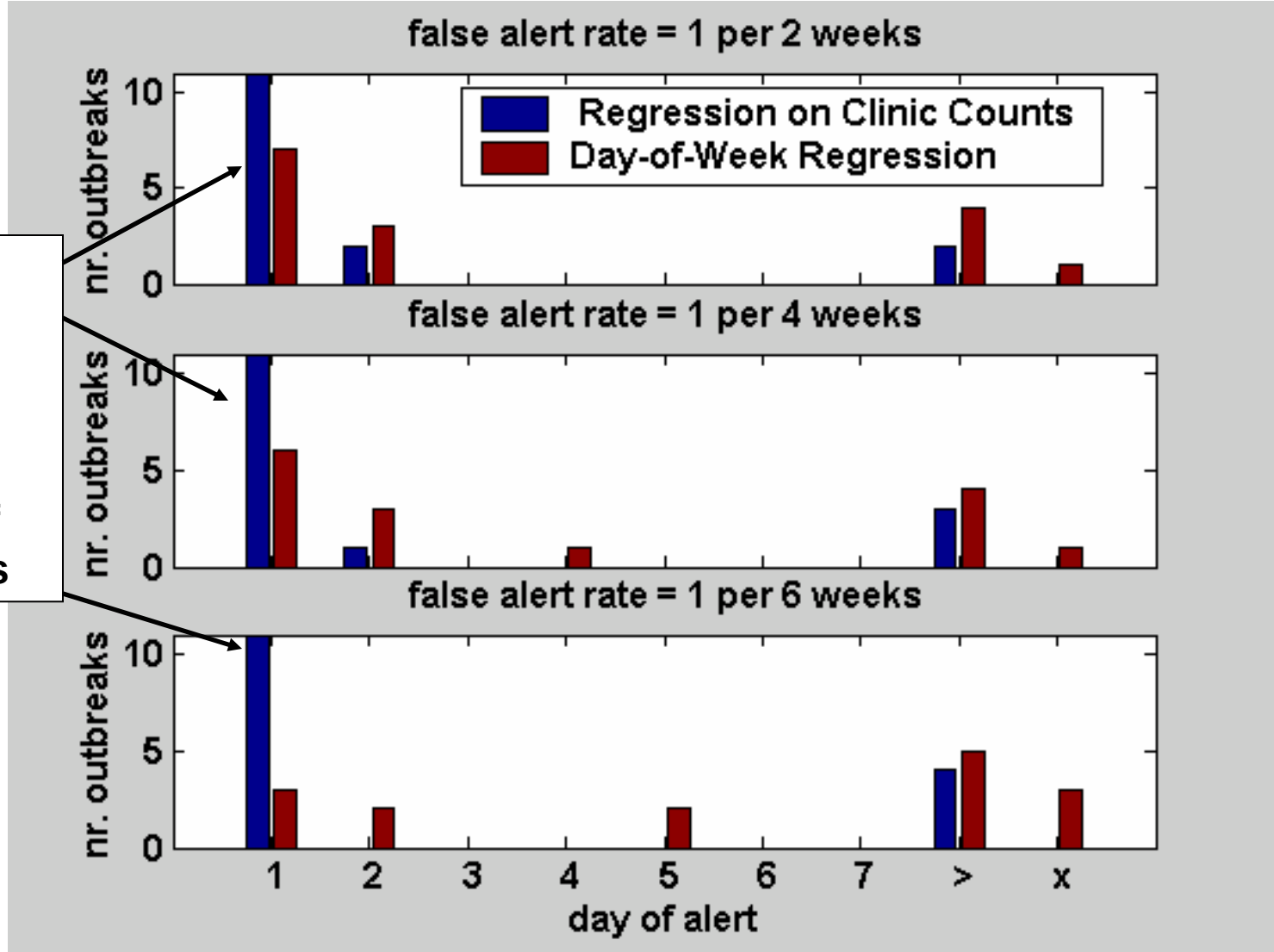


Results Using DARPA BioALIRT 2003 Evaluation Dataset--Counts only from Next-Day Reports

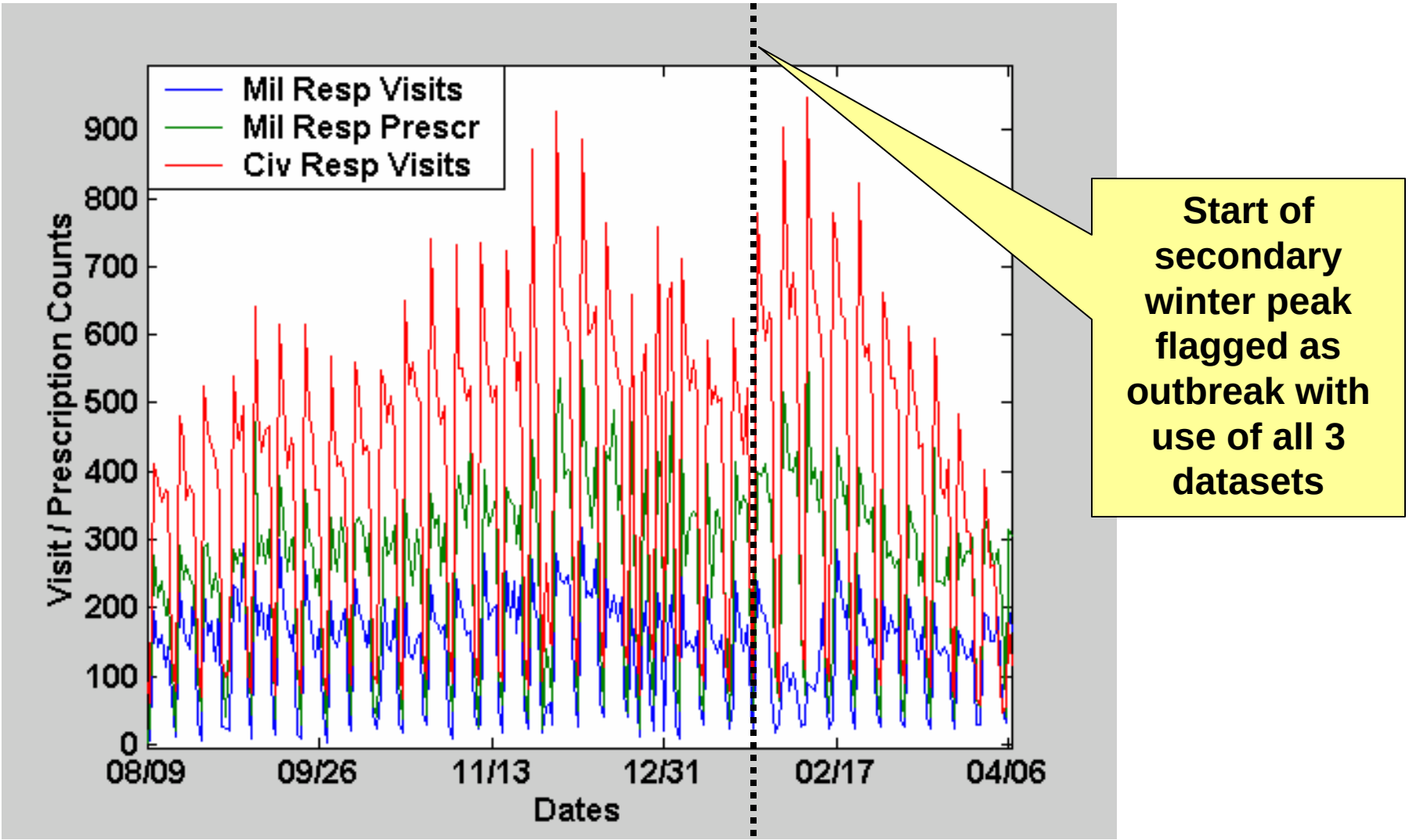
Prompt Alerting,

Improved Sensitivity (15 outbreaks)

Using Number of Reporting Clinics

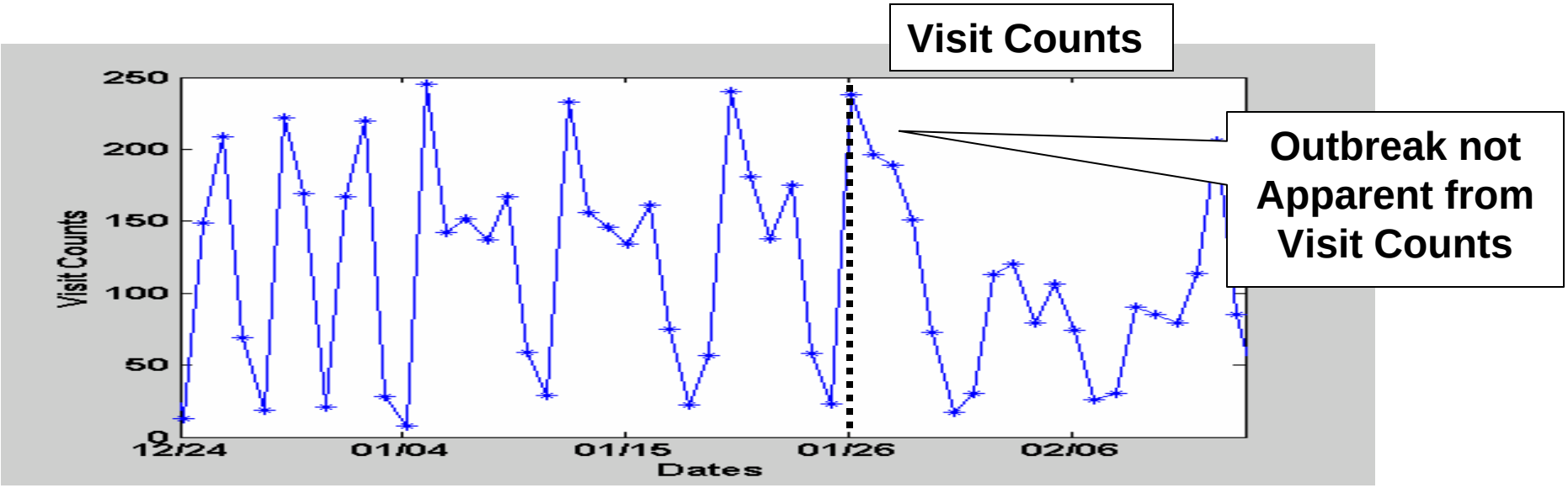


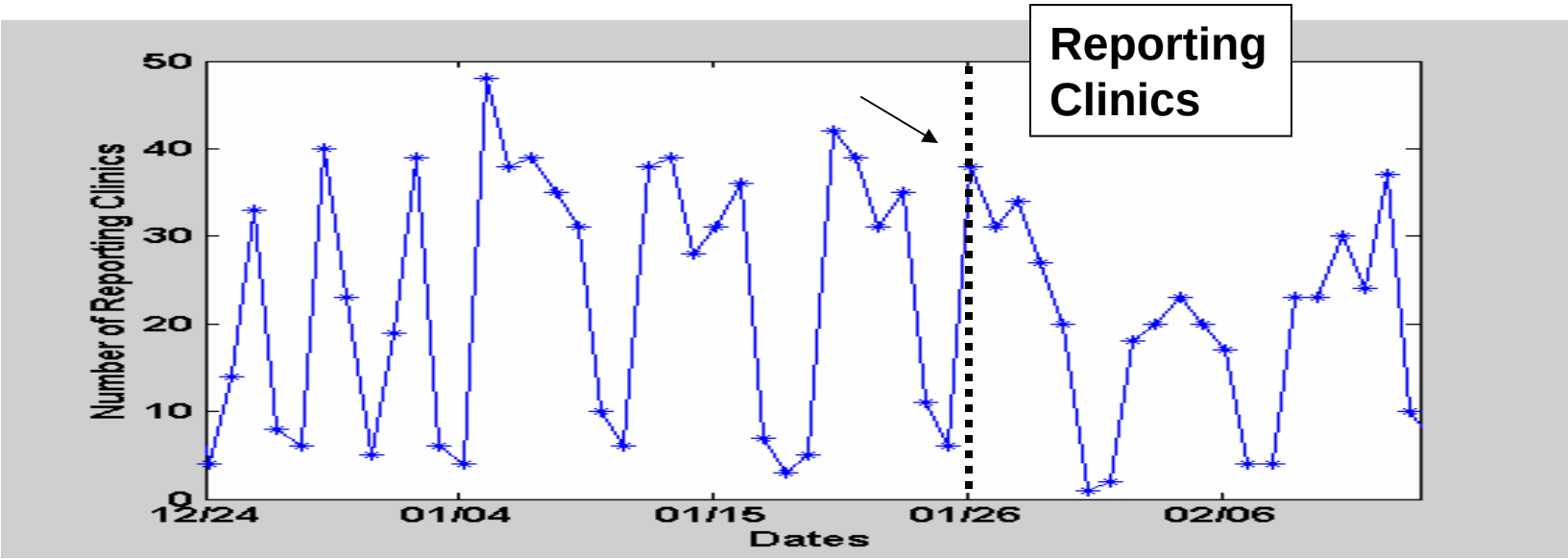
Example: Respiratory Outbreak
DARPA BioALIRT Evaluation Data
Late Winter 2003



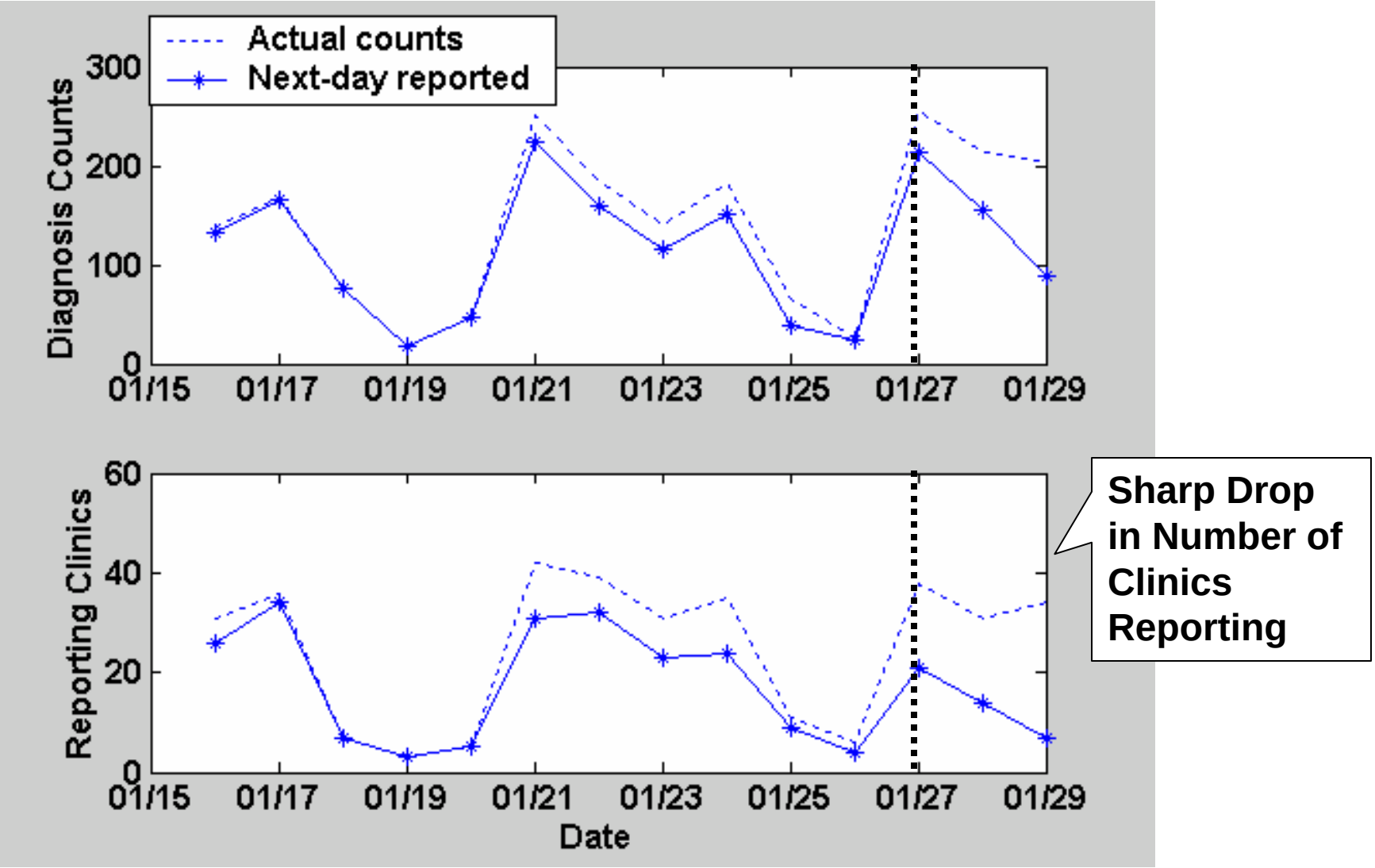
Alert from Military Visits Alone

Enabled by Including Drop in Reporting Clinics



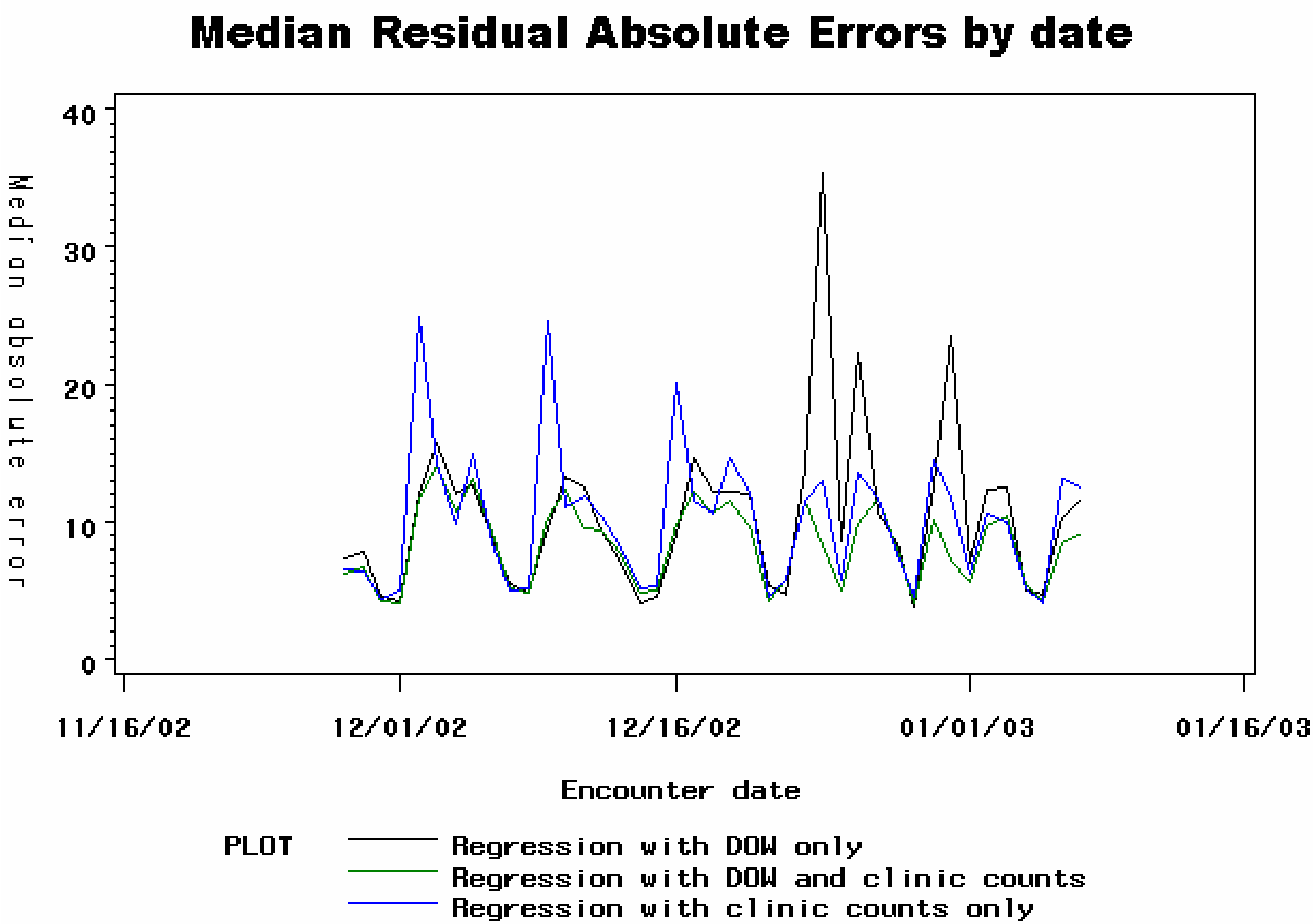


Outbreak Alert from Military Visits Alone
What the next-day algorithms saw:



Comparison of regression predictions for respiratory syndrome visits using reported vs actual counts

- One year of data: 08/01/2002 – 07/31/2003
- Counts from 100 Military Treatment Facilities with fairly “large” average number of respiratory visits(>25/day)
- Study Conditions:
  - “Next day counts”: regression using counts of all visits <u>known at surveillance time</u>, *including backfill*
  - Example, for 31Jan analysis, 28Jan counts include reports $\leq$ 3 days late
  - Actual Counts : Each day, regression using all respiratory visits regardless of report date


Median Residual Absolute Errors by date
Median absolute error
40
30
20
10
0
11/16/02
12/01/02
12/16/02
01/01/03
01/16/03
Encounter date
PLOT
Regression with DOW only
Regression with DOW and clinic counts
Regression with clinic counts only

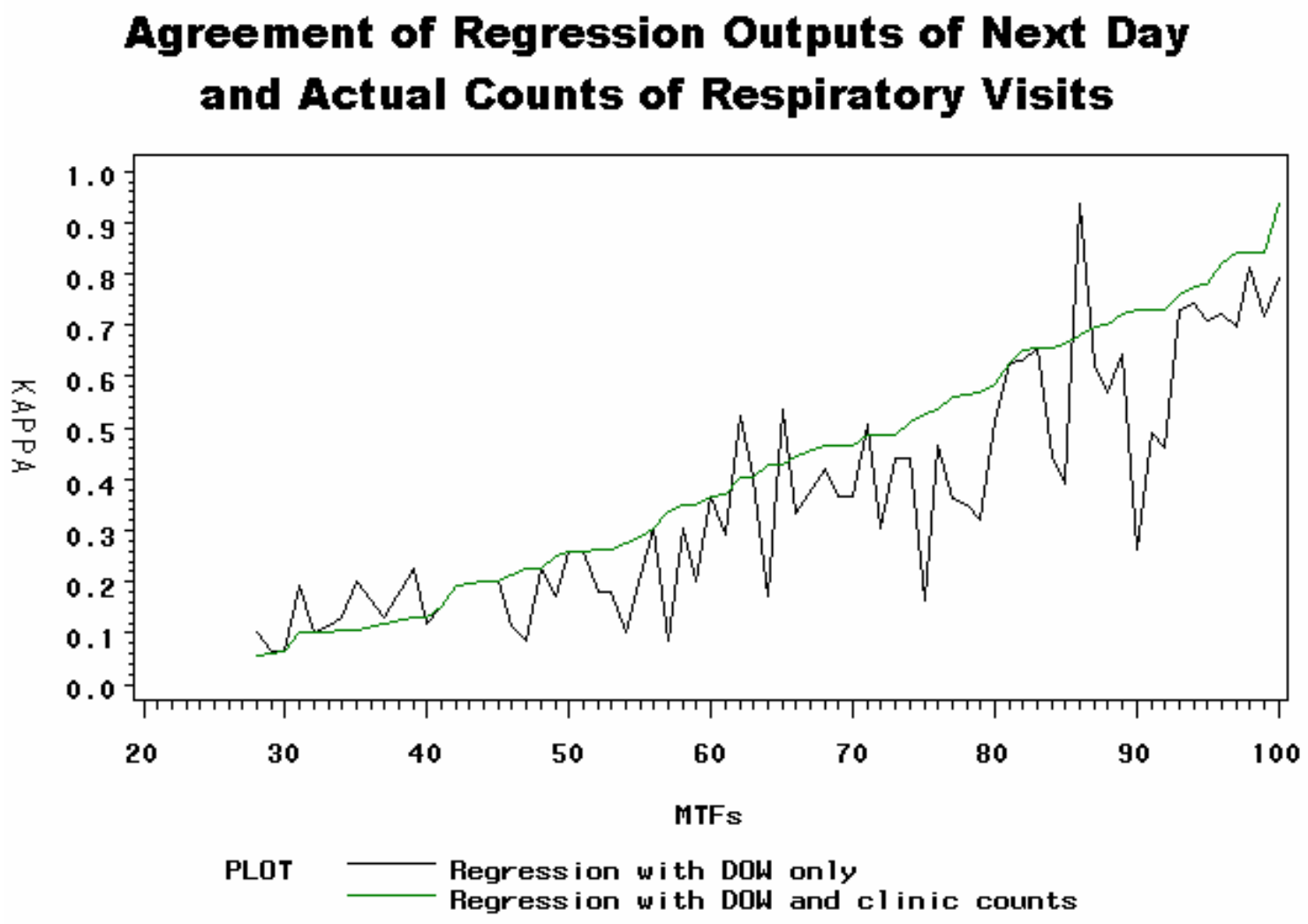
Conclusions

- Detailed standards are required for objective assessment of data sources
  - Benchmark criteria are required
- Late reporting effects measurably reduce utility of data sources for surveillance
  - Weighted kappa measures late reporting degradation
- Statistical adjustment can reduce late reporting effects
  - If provider count data are available, adjustments may be added to data modeling

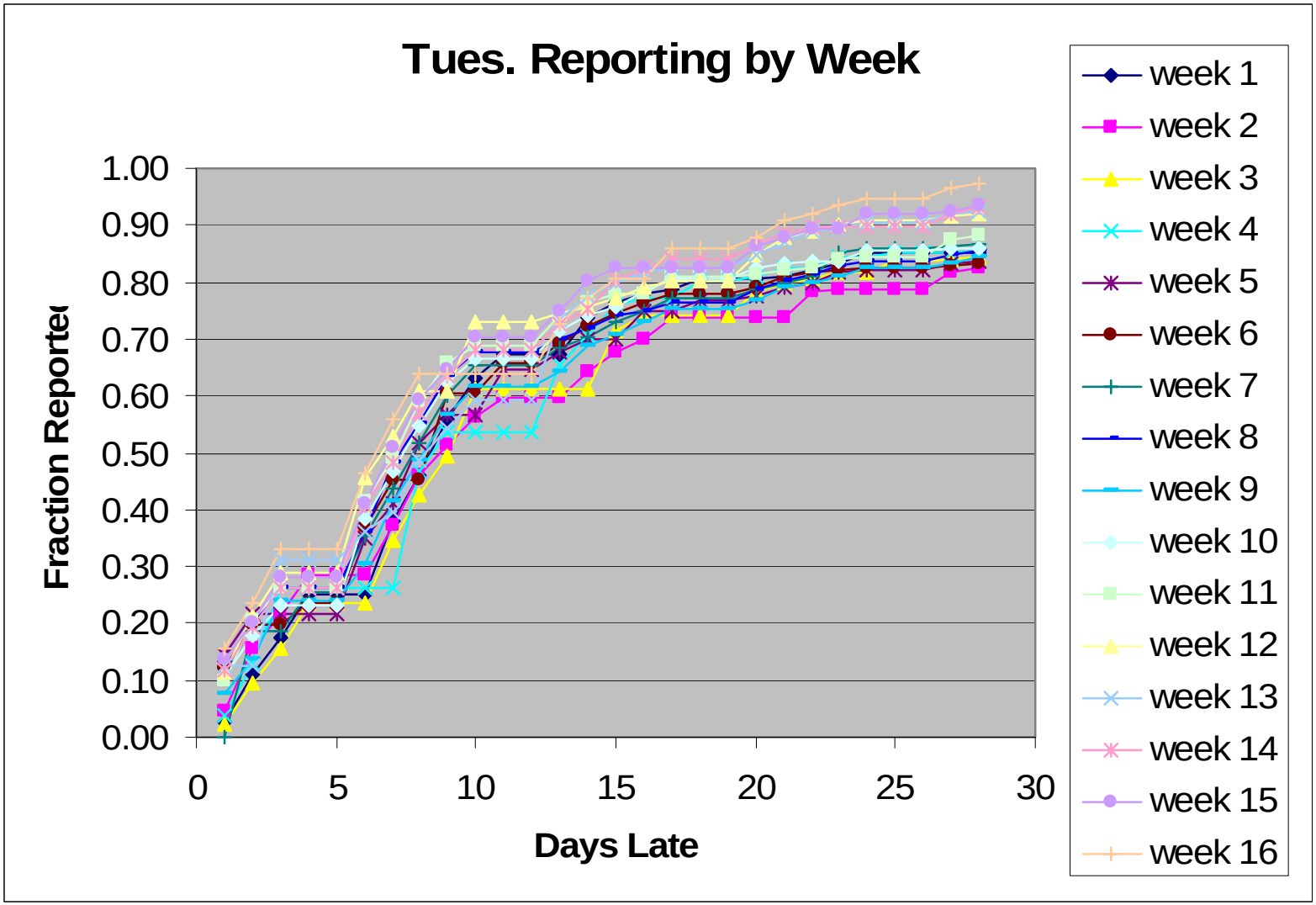
BACKUPS

Kappa Weighting

	p<0.01	p<0.05	no flag
p<0.01	1	0.75	0
p<0.05	0.75	1	0
no flag	0	0	1


Agreement of Regression Outputs of Next Day
and Actual Counts of Respiratory Visits
KAPPA
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
20
30
40
50
60
70
80
90
100
MTFs
PLOT
Regression with DOW only
Regression with DOW and clinic counts

Consistency of Reporting
Civilian #1 Respiratory Visits



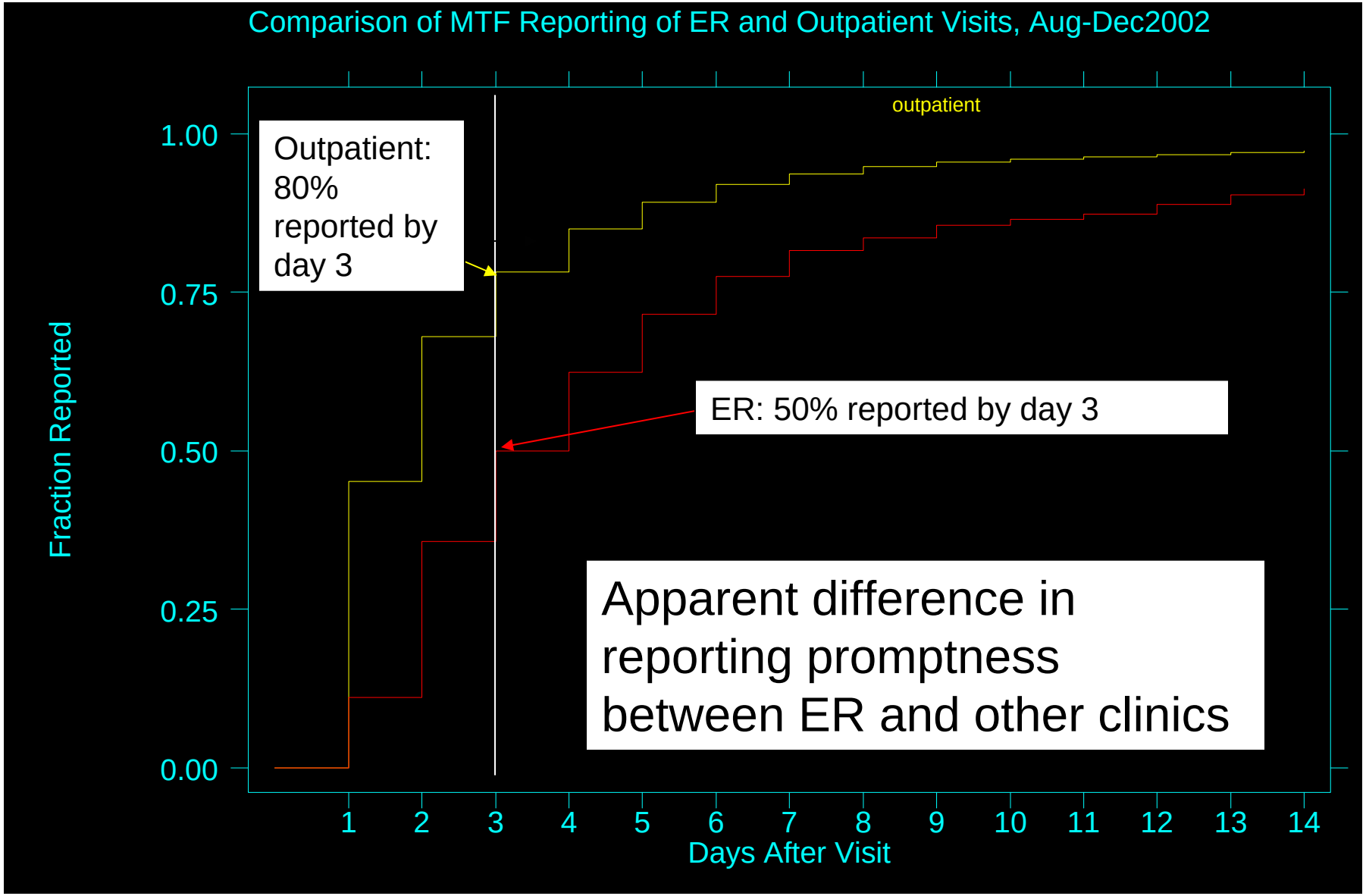
Reporting: Military Outpatient Visits

No July 4 Reporting

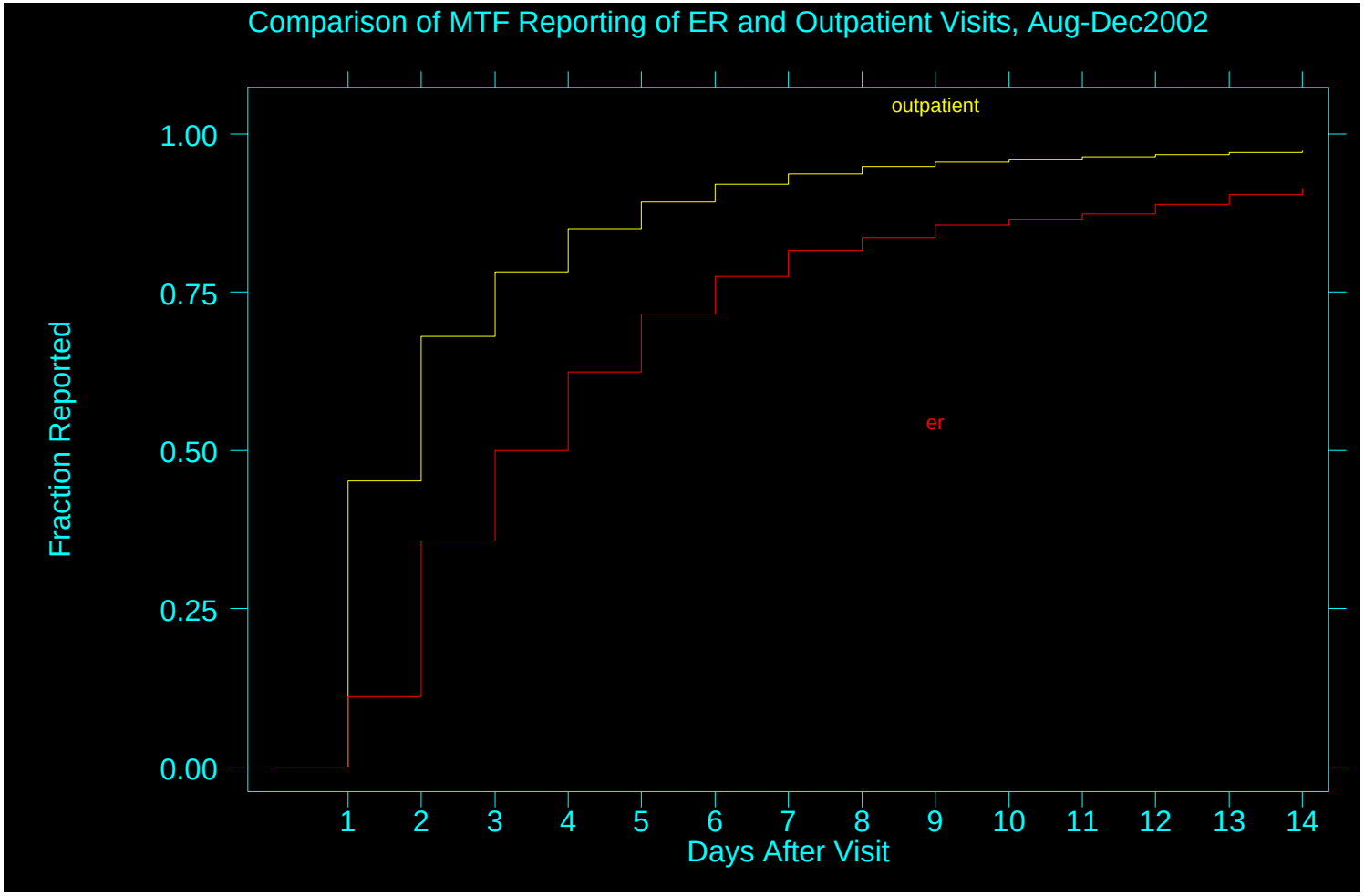
Sundays in red

Military Resp. Synd.		Reporting Lag in Days													
Sum of Visit Count	Lag														
Visit Date	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
07/01/03	0	70	29	0	1	0	0	5	21	5	0	8	0	0	0
07/02/03	0	63	0	0	1	1	10	8	2	0	1	0	0	12	2
07/03/03	0	0	5	4	0	39	12	6	2	1	1	0	9	0	3
07/04/03	0	8	7	0	14	0	0	0	1	0	0	8	0	0	0
07/05/03	0	23	2	8	4	0	0	0	0	0	1	1	6	0	8
07/06/03	0	3	18	7	1	0	0	0	0	5	2	1	0	14	0
07/07/03	0	63	38	12	2	5	0	1	2	0	3	4	40	0	0
07/08/03	0	66	26	8	10	0	2	1	0	4	1	32	3	0	0
07/09/03	0	66	21	9	2	7	7	0	5	7	36	0	0	5	0
07/10/03	0	25	30	4	2	15	4	5	3	13	0	0	4	1	4
07/11/03	0	59	3	1	18	13	7	0	24	1	0	0	7	3	0
07/12/03	0	14	0	0	5	0	0	4	5	0	1	2	3	0	0
07/13/03	0	1	0	0	8	0	8	5	0	1	0	4	0	0	0
07/14/03	0	61	16	23	8	16	4	0	12	0	5	0	0	1	0

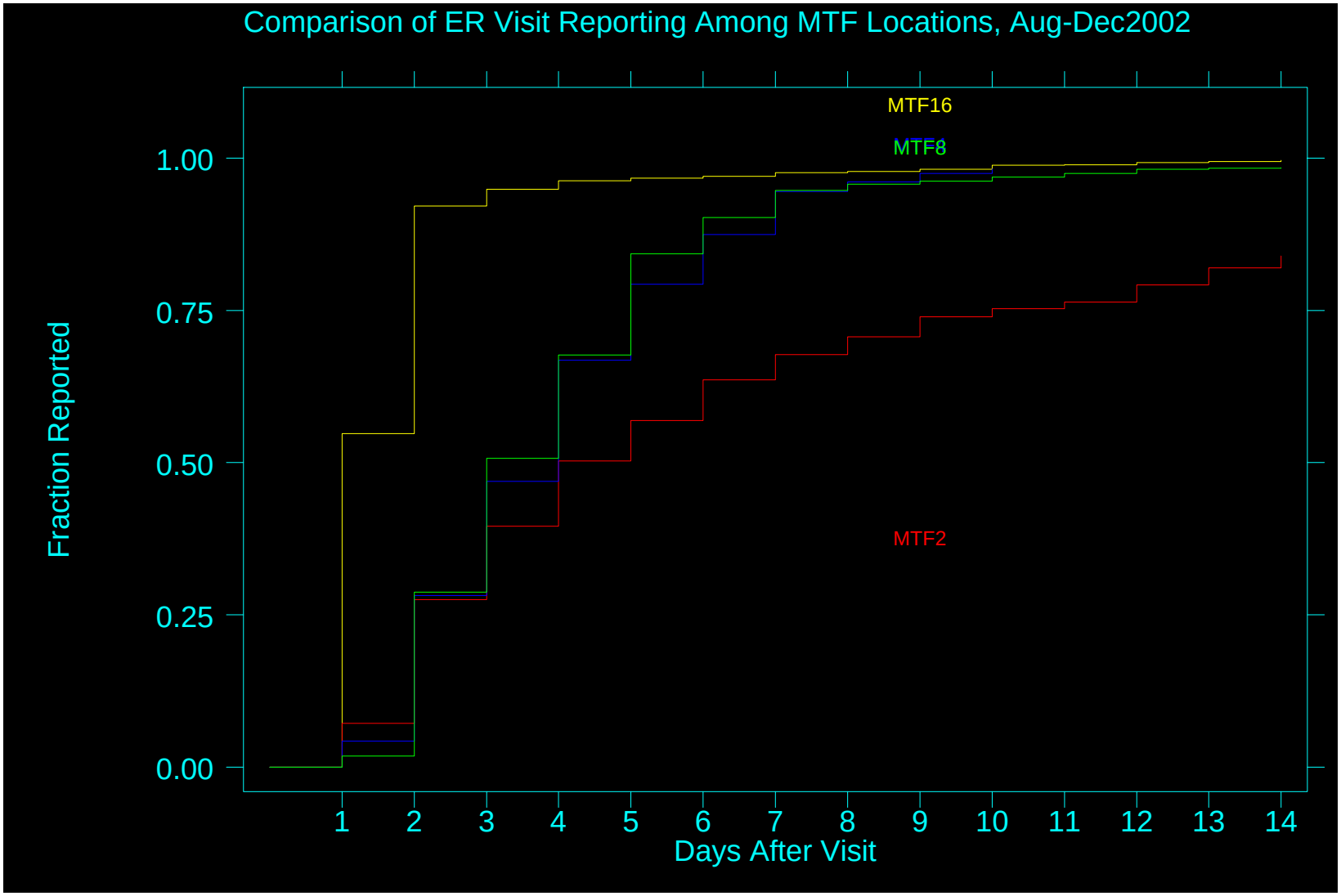
Reporting of ER/Outpatient Visits



Reporting of ER/Outpatient Visits

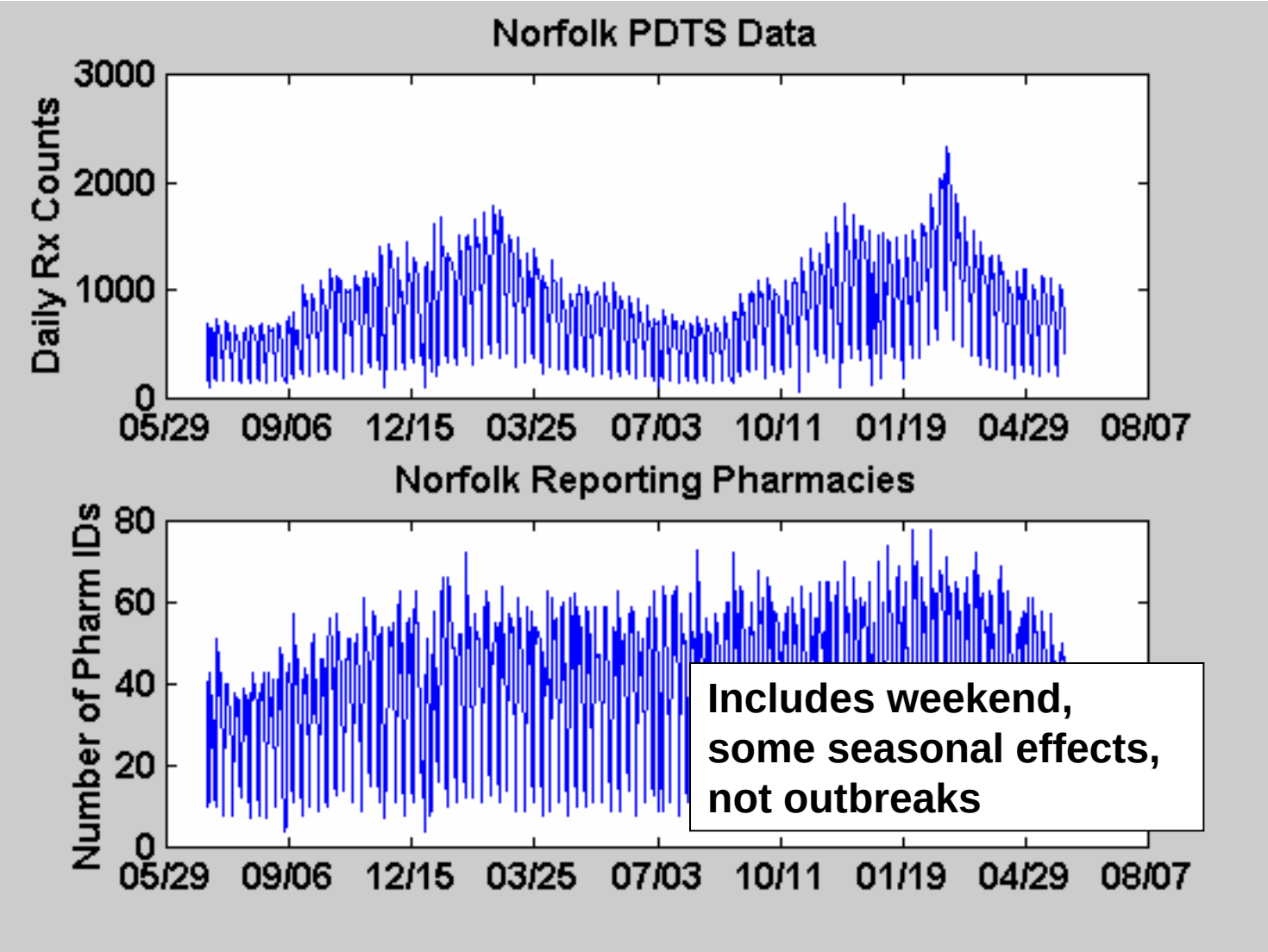


ER Visit Reporting By Location



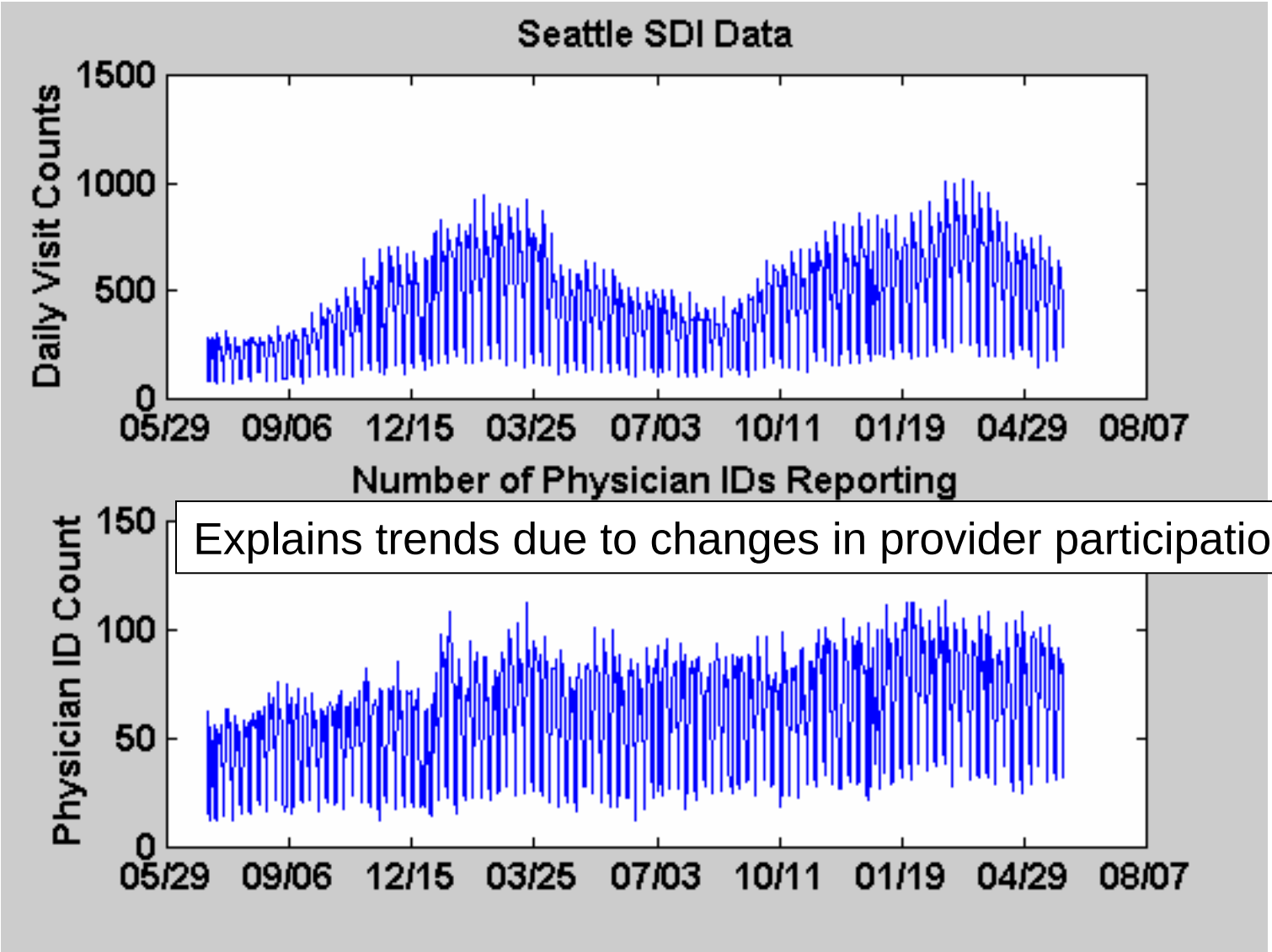
Counts of Pharmacy IDs

PDTS Data



Counts of Physician IDs

SDI Data



Explains trends due to changes in provider participation