

Syndromic Surveillance Without Denominator Data: The Space-Time Permutation Scan Statistic

Martin Kulldorff

Harvard Medical School and
Harvard Pilgrim Health Care

Farzad Mostashari, Rick Heffernan
New York City Department of Health

Jessica Hartman

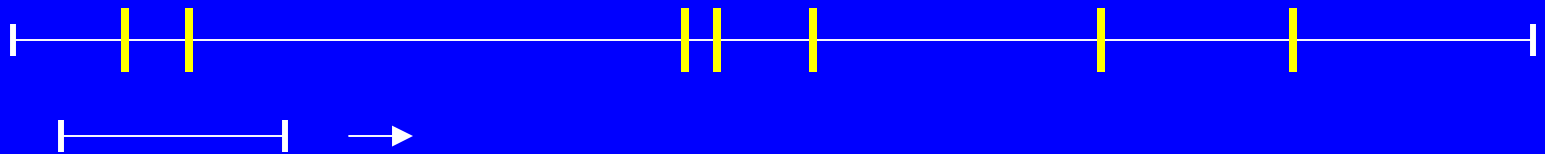
New York Academy of Medicine

Why Use a Space-Time Scan Statistic?

With disease outbreaks:

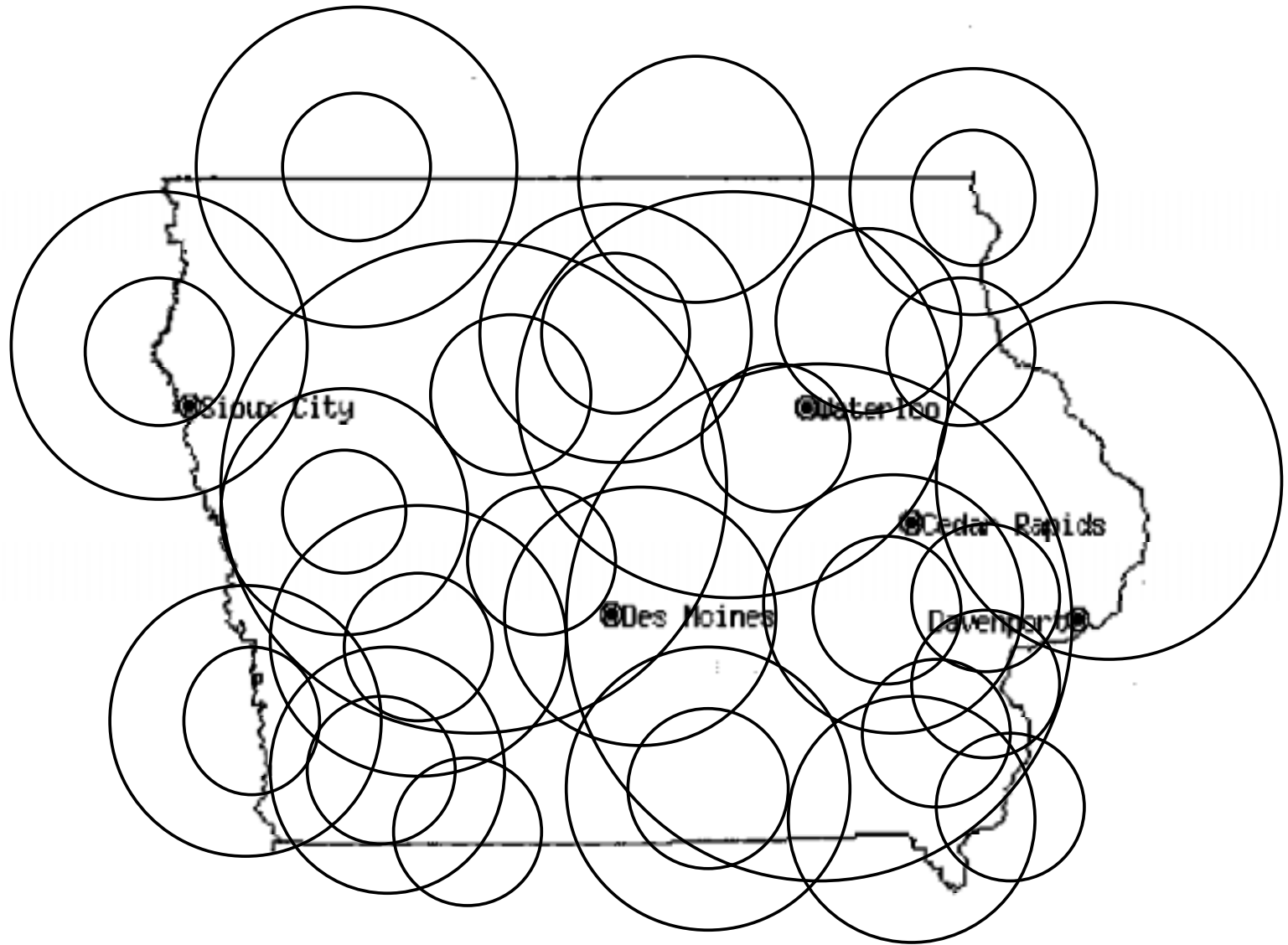
- We do not know when they will occur.
- We do not know how rapidly they will emerge.
- We do not know where they will occur.
- We do not know their geographical size.

One-Dimensional Scan Statistic



The Spatial Scan Statistic

- Create a regular or irregular grid of centroids covering the whole study region.
- Create an infinite number of circles around each centroid, with the radius varying continuously from zero up to a maximum.



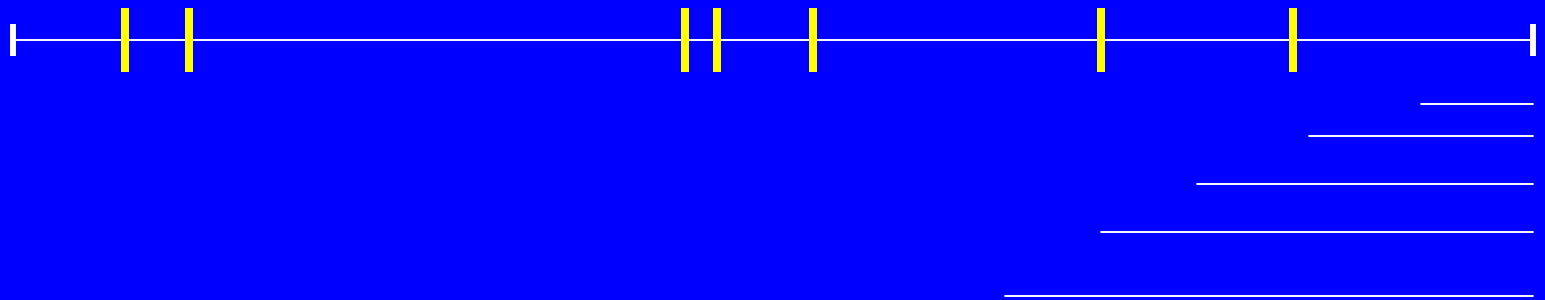
A small sample of the circles used

Space-Time Scan Statistic

Use a cylindrical window, with the circular base representing space and the height representing time.

When interested in prospective analyses for outbreak detection, use only cylinders that reach the present time.

Prospective Scan Statistic



For each cylinder:

- Obtain actual and expected number of cases inside and outside the cylinder.
- Calculate likelihood function.

Compare cylinders:

- Pick cylinder with the maximum likelihood. This is the most likely cluster.

Inference:

- Generate random replicas of the data set under the null-hypothesis of no clusters (Monte Carlo sampling).
- Compare most likely clusters in real and random data sets (Likelihood ratio test).

Space-Time Scan Statistic: Properties

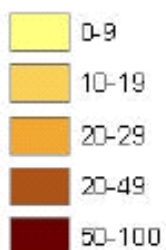
- Adjusts for inhomogeneous population density.
- Simultaneously tests for clusters of any size and any location, by using cylindrical windows with continuously variable radius.
- Accounts for multiple testing.
- Possibility to include confounding variables, such as age, sex or socio-economic variables.
- Aggregated or non-aggregated data (states, counties, census tracts, block groups, households, individuals).

Examples of Denominator Data

- Census population data
- HMO members
- All school children
- All ED visits
- All pharmacy sales

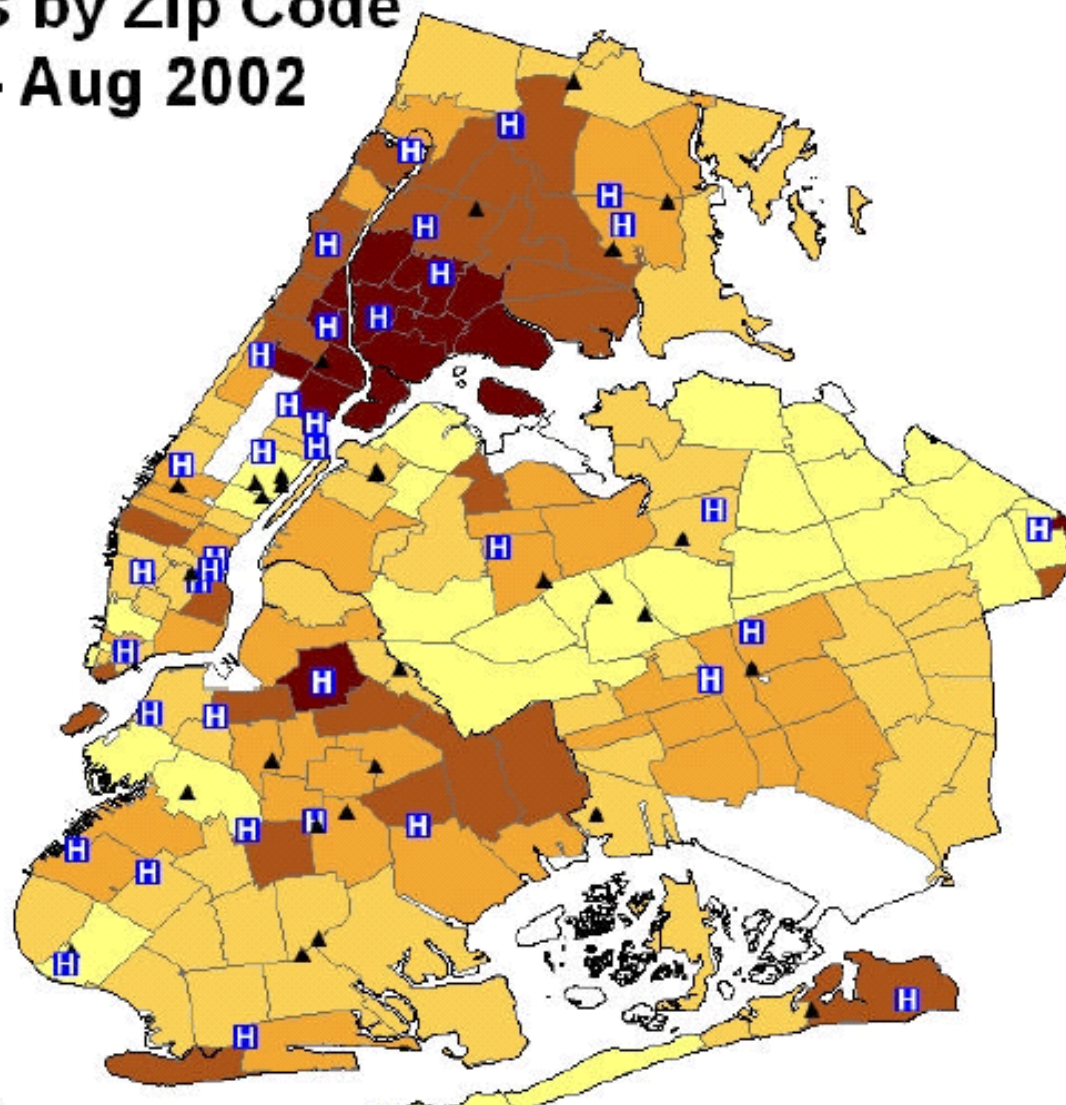
Reported ED Visit Rates by Zip Code New York City, May - Aug 2002

Monthly reported ED visits per
1000 pop (Census 2000)



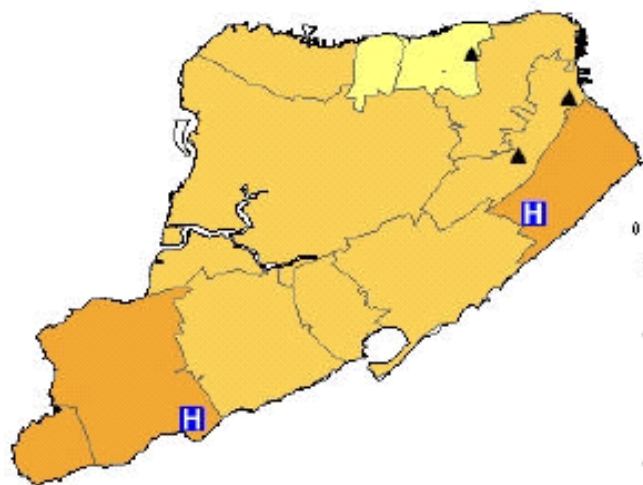
 Participating EDs

 Other EDs



•38 (56%) of 68 NYC EDs

•73% of ED visits



Adjusted Denominators

- Age
- Gender
- Sales promotions
- Temperature
- Seasonal trends
- Day-of-week variation
- Holidays
- Zip-code
- Known outbreaks
- Missing data
- Zip-code by day-of-week interaction

Potential Problem

- No denominator data available
- Needed covariates not available
- Complex denominator adjustments

Space-Time Permutation Scan Statistic

- Needs only case data
- Historical cases used to calculate the expected.
- Adjusts for purely spatial clusters
- Adjusts for purely temporal clusters

Expected

Example:

- Zip code 10029 has 2 percent of all cases in New York City, from Jan 1 to Oct 24.
- Today there are 100 cases in all of New York City.
- Today the expected in zip code 10029 is $100 \times 2\% = 2$.

Space-Time Permutation Scan Statistic

For each cylinder, calculate the expected number of cases conditioning on the marginals

$$\mu_{st} = \sum_s c_{st} \times \sum_t c_{st} / C$$

where c_{st} = # cases at time t in location s
and C = total number of cases

Space-Time Permutation Scan Statistic

For each cylinder, calculate

$$T_{st} = \left[\frac{c_{st}}{\mu_{st}} \right]^{c_{st}} \times \left[\frac{(C - c_{st})}{(C - \mu_{st})} \right]^{C - c_{st}}$$

Test statistic $T = \max_{st} T_{st}$

Statistical Inference

- Generate random replicas of the data set conditioned on the marginals, by permuting the pairs of spatial locations and times.
- Compare test statistic in real and random data sets using Monte Carlo hypothesis testing (Dwass, 1957):

$$p = \text{rank}(T_{\text{real}}) / (1 + \#\text{replicas})$$

Null Occurrence Rate

How often will a signal of the observed magnitude occur by chance?

Null Occurrence Rate = once every f/p days

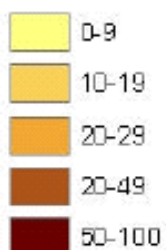
where f = days between repeated analyses
 p = p-value

NYC Emergency Department Syndromic Surveillance System

- Daily reports of visits to hospital emergency departments
- Information about nature of the visit and residential zip-code
- Daily analyses for outbreaks

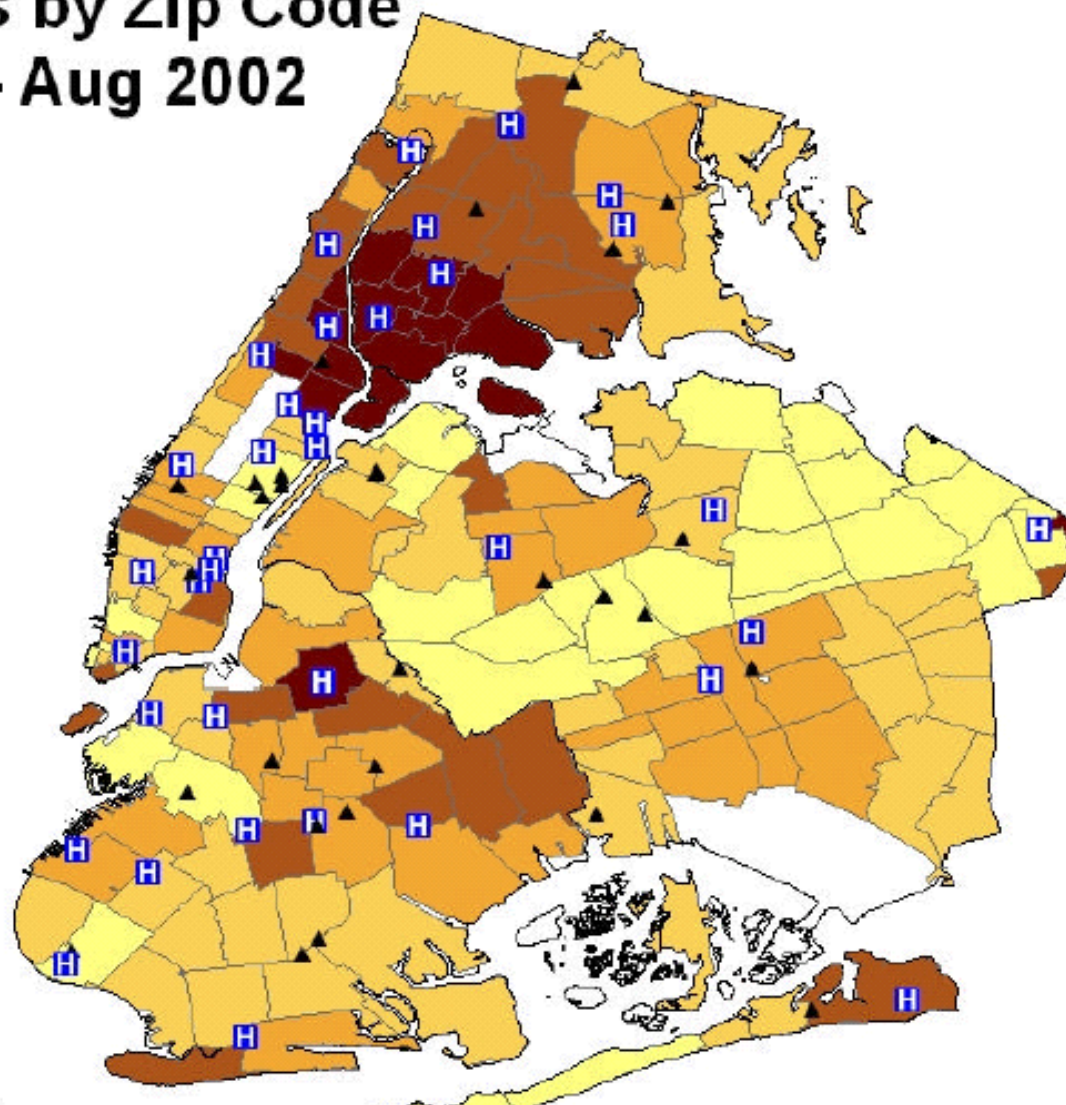
Reported ED Visit Rates by Zip Code New York City, May - Aug 2002

Monthly reported ED visits per
1000 pop (Census 2000)



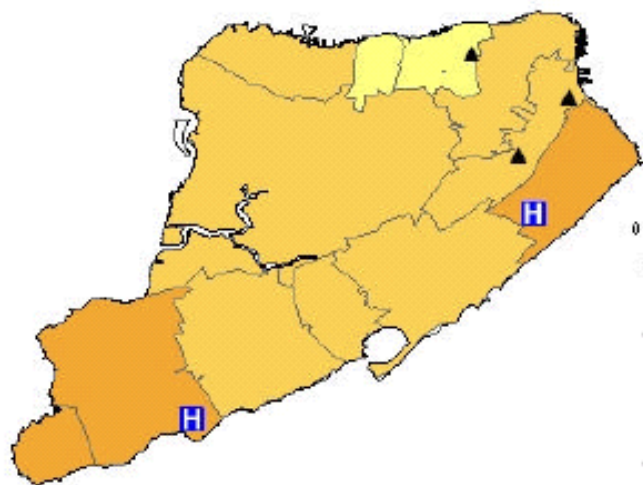
 Participating EDs

 Other EDs

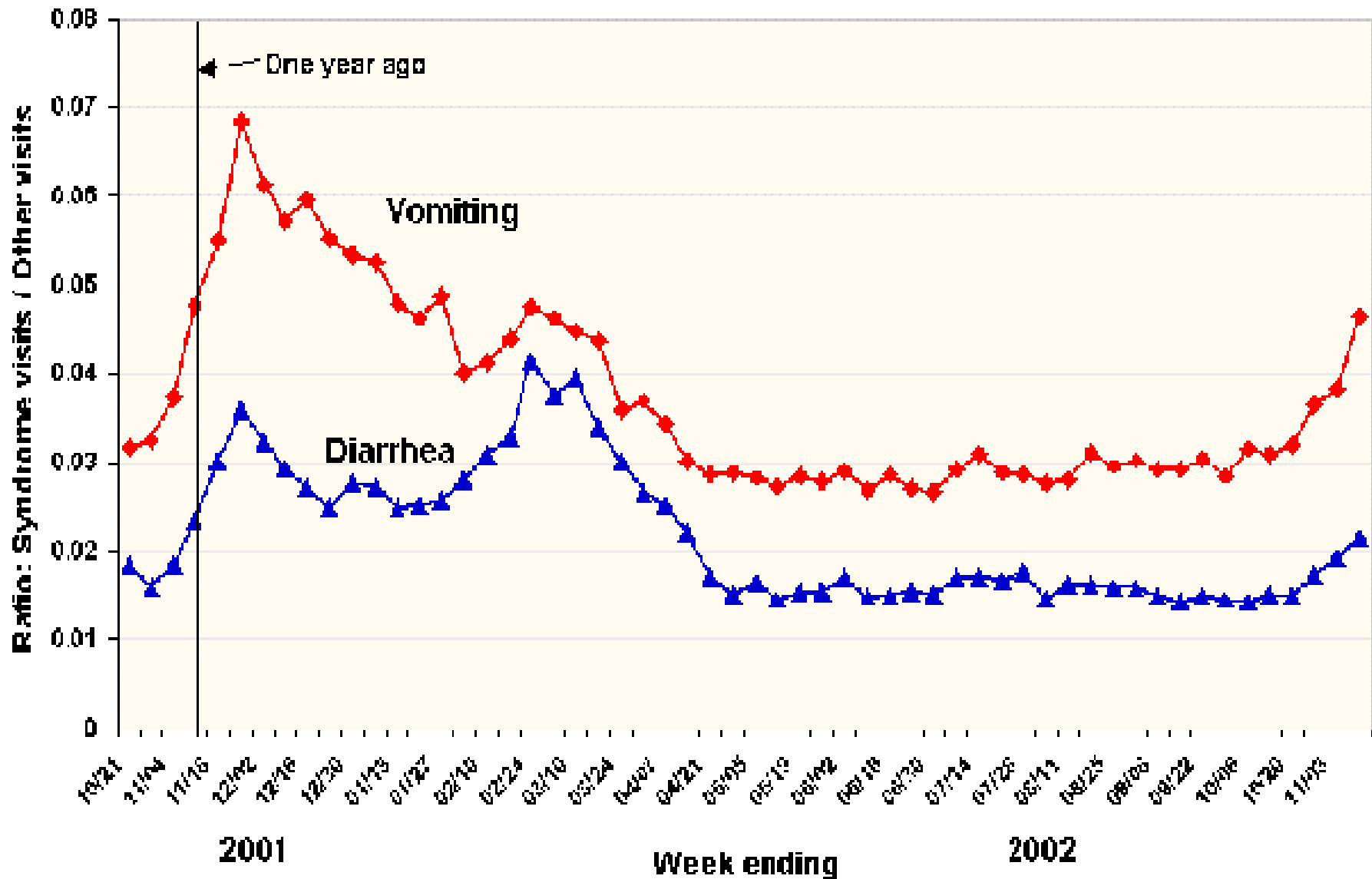


•38 (56%) of 68 NYC EDs

•73% of ED visits



Weekly Emergency Department Visits for Vomiting and Diarrhea Syndrome, New York City, All ages, Oct 2001 - Nov 2002



Data Availability

- From October 2001, until present.
- Reliable spatial data starts October 26.

Application #1

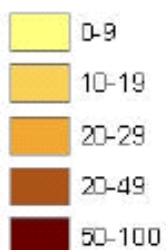
- Diarrhea, all age groups
- Data available from October 26
- Surveillance starts 2 weeks later: November 8
- Spatial window size: 0-10 kilometers
- Temporal window size: 1-7 days

Results

Day	Primary Cluster	#Days	Radius	Cases	Expected	RR	p=
Nov 8	11201+5more	2	2.2	14	5.1	2.8	0.07
Nov 9	11377 + 10 more	1	3.1	15	6.6	2.3	0.44
Nov 10	11215 + 8 more	2	3.2	21	10.0	2.2	0.14
Nov 11	11207 + 28 more	3	9.0	141	113	1.2	0.71
Nov 12	11210, 11230, 11226, 11229	1	2.7	14	5.4	2.6	0.23
Nov 13	10451	1	0	6	1.2	4.8	0.24
Nov 14	11421	1	0	6	0.8	8.0	0.009
Nov 15	11385, 11379, 11421, 11208	3	2.7	36	18.4	2.0	0.04
Nov 16	10306 + 7 more	2	7.9	9	2.0	4.4	0.04
Nov 17	11414, 11421, 11208 + 6 more	5	4.5	67	42.7	1.6	0.08
Nov 18	11218 + 7 more	1	3.0	19	7.1	2.7	0.02
Nov 19	11218 + 6 more	2	2.7	25	12.0	2.1	0.17
Nov 20	11225 + 4 more	3	2.0	46	27.9	1.6	0.25
Nov 21	11225 + 6 more	4	2.3	66	44.0	1.5	0.35

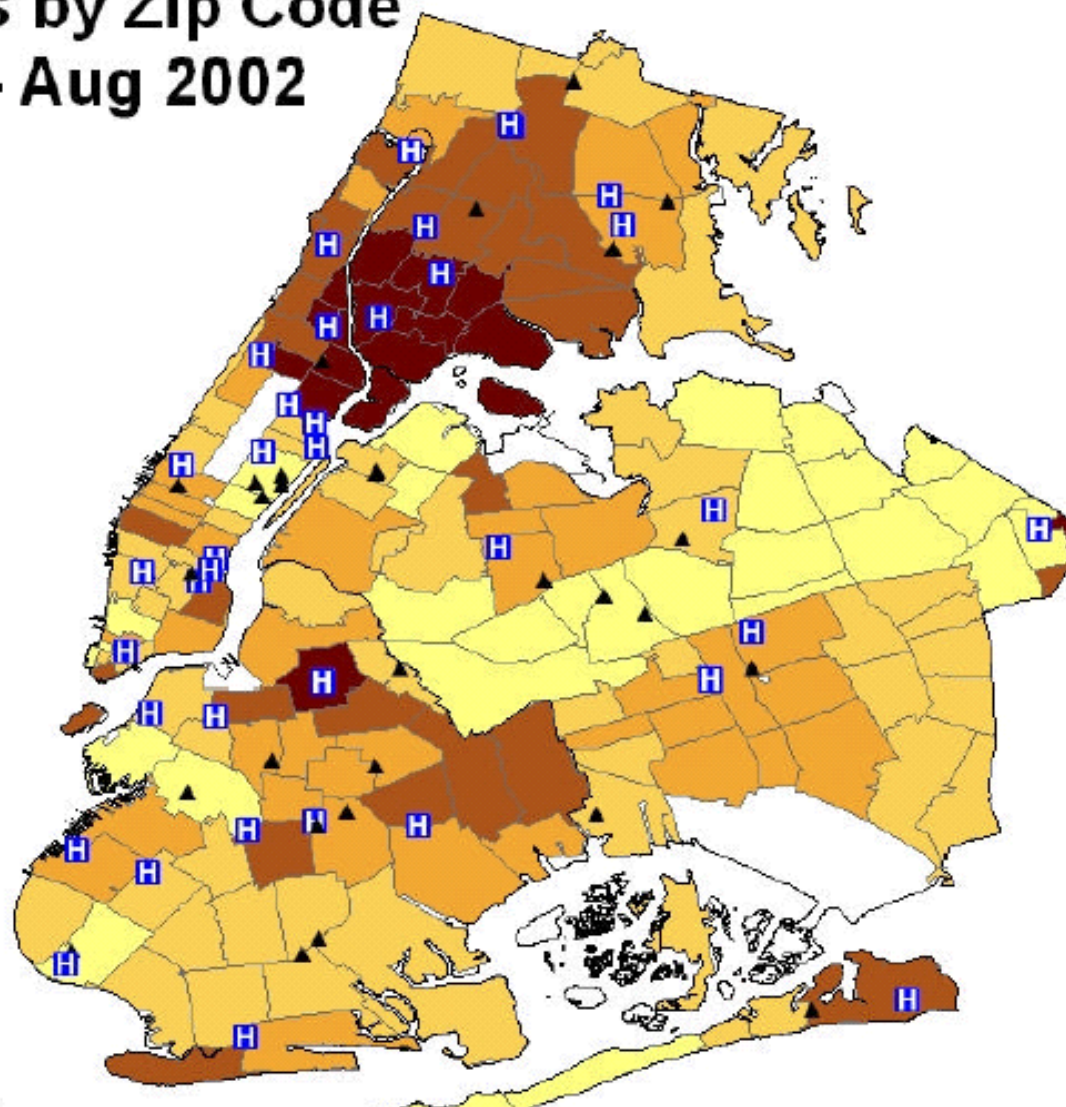
Reported ED Visit Rates by Zip Code New York City, May - Aug 2002

Monthly reported ED visits per
1000 pop (Census 2000)



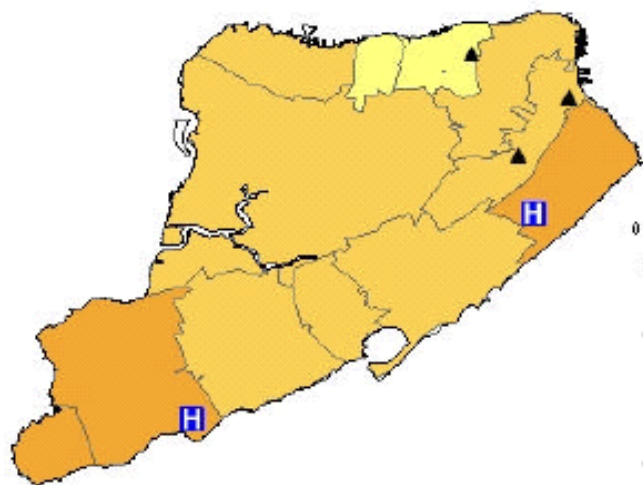
 Participating EDs

 Other EDs



•38 (56%) of 68 NYC EDs

•73% of ED visits



Outbreak Investigation

- School Children
- ~ 150 total cases
- 19 vomited at school after lunch
- 18 visited hospital emergency departments
- Calicivirus

What If ?

- Hospital A tends to have many ED visits on weekdays
- Hospital B tends to have many ED visits on weekends.

Adjustment for Day-of-Week/Space Interaction

Problem: Weekly variation in rates may differ in different areas

Solution: Randomize stratified by day-of-week

Warning: Less amount of baseline data may create problems

Missing Data

Three Options:

- Remove geographical area (or data source) for all days.
- Remove day for all geographical areas
- If day-of-week is a covariate, remove geographical area only for selected days of the week.

Year Long Run

- Daily analyses of ER diarrhea visits
- From November 2001 to November 2002
- Thirty days of data used in each run.
- Max # days = 7, Max radius = 5km
- Adjusted for spatial / day-of-week interaction.
- Adjusted for missing data, using all three methods at different times.
- Cut-off: One false signal per year ($p < 0.0028$)

Results: Two Signals

February 9, 2002:

Bronx, 15 zip-codes, 2 days

63 observed, 34.7 expected

$p=0.0001$

null occurrence rate = once every 27.5. years

March 7, 2002:

Northern Manhattan, 8 zip-codes, 2 days

63 observed, 37.3 expected

$p=0.0027$

null occurrence rate = once every 12.2 months

Limitations

- Space-time clusters may occur for other reasons than disease outbreaks
- Automated detection systems does not replace the observant eyes of physicians and other health workers.
- Epidemiological investigations by public health department are needed to confirm or dismiss the signals.

Conclusions

- The space-time permutation scan statistic can serve as an important tool in prospective systematic time-periodic geographical surveillance for the early detection of disease outbreaks.
- Only case data are needed.

SaTScan v4.0 Software

<http://www.satscan.org/>

- Temporal, Spatial, Space-Time Analyses
- Poisson, Bernoulli, Space-Time Permutation
- High or Low Risk Areas

Sample Speed

Purely spatial scan statistic, 10,000 locations,
10% maximum window and 999 Monte Carlo
replications:

18 minutes

Clarification

The “fast” algorithm developed for the spatial scan statistic and mentioned by Mike Wagner was compared to a “standard” algorithm developed in Pittsburgh. The speed has never been compared to the algorithm in the SaTScan software. (Designed for different types of spatial data.)