

A Space-Time Permutation Scan Statistic for the Early Detection of Disease Outbreaks

Martin Kulldorff
Harvard University Medical School and
Harvard Pilgrim Health Care Institute

S-GEM Workshop, Stockholm, 2009

Importance of Early Disease Outbreak Detection

- Eliminate health hazards
- Warn about risk factors
- Earlier diagnosis of new cases
- Quarantine cases
- Scientific research concerning treatments, vaccines, etc.
- Early detection is especially critical for infectious diseases

Prospective Disease Surveillance Data Sources

- Disease registries
- Reportable diseases
- Electronic health services records
- Pharmacy sales
- etc

Purely Temporal Methods

Farrington CP, Andrews NJ, Beale AD, Catchpole MA (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. *J R Stat Soc A Stat Soc* 159: 547–563.

Hutwagner LC, Maloney EK, Bean NH, Slutsker L, Martin SM (1997) Using laboratory-based surveillance data for prevention: An algorithm for detecting salmonella outbreaks. *Emerg Infect Dis* 3: 395–400.

Nobre FF, Stroup DF (1994) A monitoring system to detect changes in public health surveillance data. *Int J Epidemiol* 23: 408–418.

Reis B, Mandl K (2003) Time series modeling for syndromic surveillance. *BMC Med Inform Decis Mak* 3: 2.

Three Important Issues

- An outbreak may start locally.
- Purely temporal methods can be used simultaneously for multiple geographical areas, but that leads to multiple testing.
- Disease outbreaks may not conform to the pre-specified geographical areas.

Why Use a Scan Statistic?

With disease outbreaks:

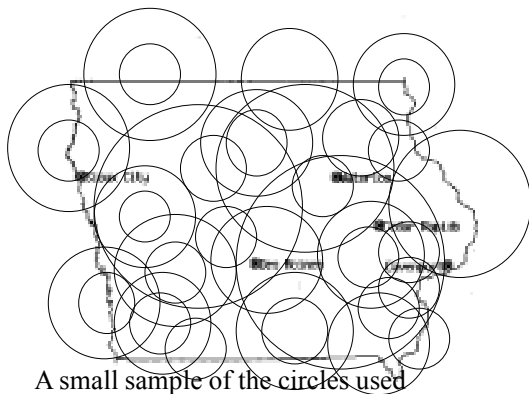
- We do not know where they will occur.
- We do not know their geographical size.
- We do not know when they will occur.
- We do not know how rapidly they will emerge.

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 anywhere from zero up to a maximum so that at most 50 percent of the population is included.



A small sample of the circles used

For each circle:

- Obtain actual and expected number of cases inside and outside the circle.
- Calculate Likelihood Function.

Compare Circles:

- Pick circle with highest likelihood function as 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).

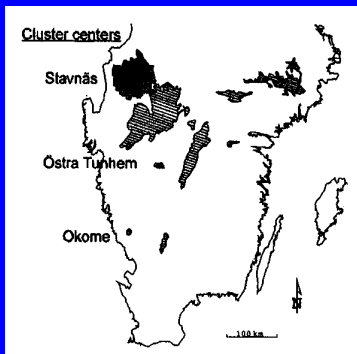
Childhood Leukemia in Sweden

Ulf Hjalmars, Martin Kulldorff
Göran Gustafsson, Neville Nagarwalla

Leukemia Incidence Data

- Acute leukemia
- Children, age 0-15 years
- Years 1973-1993
- 1523 cases
- 2577 parishes
- Denominator: 1,703,235 children, based on an average for years 1976,1982 and 1988.

Three Most Likely Clusters



Three Most Likely Clusters

	obs	exp	pop	p=
Okome	3	0.1	133	0.70
Ö Tunhem	5	0.6	695	0.91
Stavnäs	20	9.9	10380	0.99

Conclusions

- No evidence of any childhood cancer clusters in Sweden
- A “leukemia cluster” in Åstorp that received media attention in 1981 was detected, but it was not among the top three clusters nor statistically significant.

Spatial Scan Statistic: Properties

- Adjusts for inhomogeneous population density.
- Simultaneously tests for clusters of any size and any location, by using circular 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).

Space-Time Scan Statistic

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

We will only consider cylinders that reach the present time.

For each cylinder:

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

Compare Cylinders:

- Pick cylinder with highest likelihood function as 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).

For each cylinder:

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

Compare Cylinders:

- Pick cylinder with highest likelihood function as 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 Permutation Scan Statistic

1. 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

2. For each cylinder, calculate

$$T_{st} = [c_{st} / \mu_{st}]^{c_{st}} \times [(C - c_{st}) / (C - \mu_{st})]^{C - c_{st}} \quad \text{if } c_{st} > \mu_{st}$$

= 1, otherwise

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

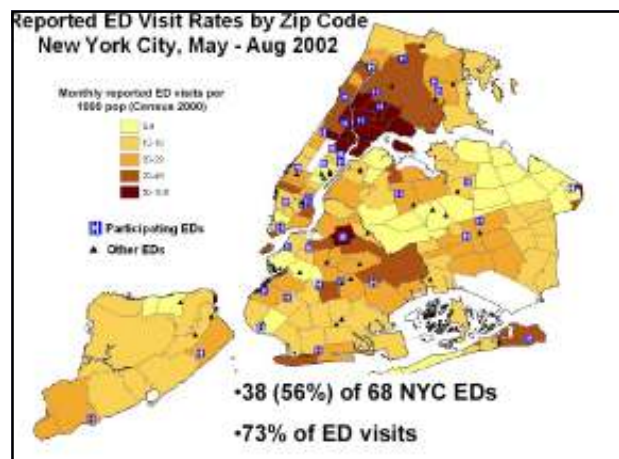
Space-Time Permutation Scan Statistic

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

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

Space-Time Permutation Scan Statistic: Properties

- Adjusts for purely geographical clusters.
- Adjusts for purely temporal clusters.
- Simultaneously tests for outbreaks of any size at any location, by using a cylindrical windows with variable radius and height.
- Accounts for multiple testing.
- Aggregated or non-aggregated data (counties, zip-code areas, census tracts, individuals, etc).



Let's Try It!

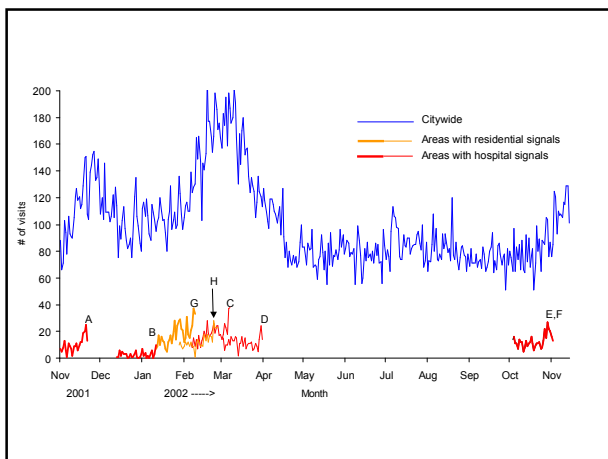
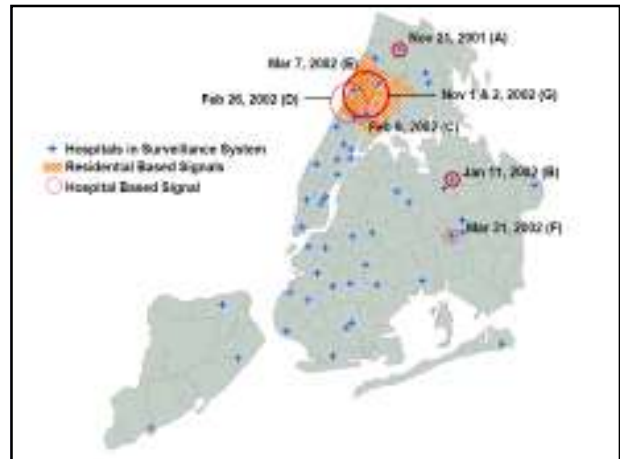
- Historic data, Nov 15, 2001 – Nov 14, 2002
- Diarrhea, all age groups
- Use last 30 days of data.
- Temporal window size: 1-7 days
- Spatial window size: 0-5 kilometers
- Residential zip code and hospital coordinates

Results: Hospital Analyses

Date	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
A Nov 21	6	1	101	73.6	1.4	0.0008	1 / 3.4 years
B Jan 11	1	1	10	2.3	4.4	0.0007	1 / 3.9 years
C Feb 26	4	2	97	66.9	1.4	0.0018	1 / 1.5 years
D Mar 31	2	1	38	19.2	2.0	0.0017	1 / 1.6 years
E Nov 1	6	3	122	86.6	1.4	0.0017	1 / 1.6 years
F Nov 2	7	3	135	98.3	1.4	0.0008	1 / 3.4 years

Results: Residential Analyses

Date	#days	#zips	#cases	#exp	RR	p=	recurrence interval
G Feb 9	2	15	63	34.7	1.8	0.0005	1 / 5.5 years
H Mar 7	2	8	63	37.3	1.7	0.0027	1 / 1.0 years



Real-Time Daily Analyses

- Starting November 1, 2003.
- Respiratory, Fever/Flu, Diarrhea, (+Vomiting)
- Hospital (and Residential) Analyses
- Spatial window size: 0-5 kilometers
- Temporal window size: 1-7 days

Real-Time Results, Nov 24, 2003: Hospital Analysis

Syndrome	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
Respiratory	2	3	80	57.4	1.4	0.13	every 8 days
Fever/Flu	3	1	24	14.8	1.6	0.68	every day
Diarrhea	2	4	18	8.2	2.2	0.04	every 26 days

Real-Time Results, Nov 25, 2003: Hospital Analysis

Syndrome	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
Respiratory	7	1	45	30.4	1.5	0.46	every 2 days
Fever/Flu	1	5	50	31.5	1.6	0.04	every 23 days
Diarrhea	3	4	22	11.5	1.9	0.17	every 6 days

Real-Time Results, Nov 26, 2003: Hospital Analysis

Syndrome	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
Respiratory	5	2	233	199.4	1.1	0.63	every 2 days
Fever/Flu	7	7	299	252.1	1.2	0.05	every 22 days
Diarrhea	4	4	23	12.6	1.8	0.22	every 5 days

Real-Time Results, Nov 27, 2003: Hospital Analysis

Syndrome	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
Respiratory	1	4	41	26.9	1.5	0.45	every 2 days
Fever/Flu	6	4	181	142.9	1.3	0.03	every 36 days
Diarrhea	5	3	29	14.1	1.7	0.50	every 2 days

Real-Time Results, Nov 28, 2003: Hospital Analysis

Syndrome	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
Respiratory	2	4	98	78.8	1.2	0.82	every day
Fever/Flu	7	5	228	178.0	1.3	0.001	every 1000 days
Diarrhea	6	3	29	17.5	1.5	0.26	every 4 days

Real-Time Results, Nov 29, 2003: Hospital Analysis

Syndrome	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
Respiratory	7	2	146	123.6	1.2	0.95	every day
Fever/Flu	7	4	253	195.7	1.3	0.001	every 1000 days
Diarrhea	7	4	44	29.4	1.5	0.21	every 5 days

Real-Time Results, Nov 30, 2003: Hospital Analysis

Syndrome	#days	#hosp	#cases	#exp	RR	p=	recurrence interval
Respiratory	1	1	19	10.7	1.8	0.69	every day
Fever/Flu	6	9	429	364.1	1.2	0.002	every 500 days
Diarrhea	1	5	12	4.4	2.7	0.06	every 17 days

Summary

Four strong diarrhea signals:

- Two were early signals for city-wide outbreaks likely due to norovirus.
- One was an early signal for a city-wide children outbreak, likely due to rotavirus.
- One small outbreak of unknown etiology.

Three medium strength diarrhea signals:

- All during the rotavirus outbreak, possibly due to a shift in the geographical epicenter

One real-time fever/flu signal, coinciding with the start of the flu season.

Shigella Surveillance in Argentina, 7/2006-6/2007

John Stelling, Katherine Yih, Martin Kulldorff
Harvard University Medical School

Marcelo Galas, Alejandra Corso,
Ezequiel Tuduri Franco
ANLIS "Dr. Carlos G. Malbran"
for the WHONET-Argentina Network

Shigella Surveillance in Argentina, 7/2006-6/2007

An evaluation of the utility of space-time scan statistics for the detection of outbreaks, using the WHONET-Argentina data.

Mimicking a daily prospective surveillance system, using historical data from July 2005 to June 2007.

Based on the evaluation, it may be possible to implement a real-time prospective surveillance system using daily or weekly data.

Analysis Specifications

Method: Space-Time Permutation Scan Statistic

Disease: All *Shigella* spp.

Surveillance Period: July 2006 to June 2007

Baseline Data: One year.

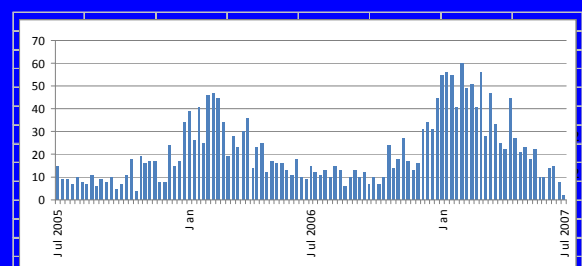
Temporal window size: 1-30 days

Spatial window size: 0-50% of all cases

Number of hospitals: 22

Minimum recurrence interval: 1 year

Purely Spatial and Purely Temporal Adjustment



Signal #1

Date: Nov 14, 2006
 Size: 1 hospital
 Length: 5 days
 Recurrence Interval:
 2.1 years
 Observed Cases: 5
 Expected cases: 0.41
 Relative Risk: 12.2



Signal #2

Date: Nov 17, 2006
 Size: 1 hospital
 Length: 2 days
 Recurrence Interval:
 9.1 years
 Observed Cases: 6
 Expected cases: 0.58
 Relative Risk: 10.3



Organisms and Resistance

Date	Org.	Resistance	AMP	SXT	CIP	NIT	NAL	Age	Sex
11/16	Shig. sp.	AMP	6		40	23	28	5	F
11/16	Shig. sp.	AMP	6		36	21	28	1	M
11/16	Shig. sp.	AMP	6		32	22	27	13	F
11/16	Shig. sp.	AMP	8		32	22	28	3	F
11/17	Shig. sp.	AMP	6			22	28	1	M
11/17	Shig. sp.	AMP	6			21	26	2	M

Signals #3a,b

Date: Dec 12, 2006
 Size: 1 hospital
 Length: 19 days
 Recurrence Interval:
 2.3 years
 Observed Cases: 5
 Expected cases: 0.41
 Relative Risk: 12.2

Date: Dec 13, 2006
 Size: 1 hospital
 Length: 20 days
 Recurrence Interval:
 13.7 years
 Observed Cases: 6
 Expected cases: 0.52
 Relative Risk: 11.5

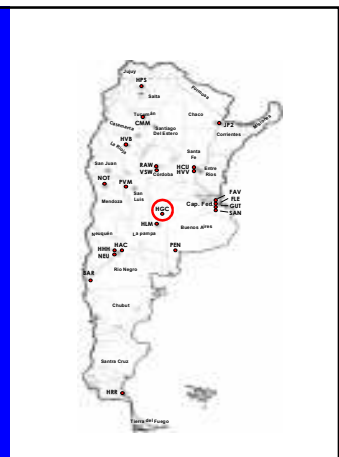


Organisms and Resistance

Date	Clone	Resistance	AMP	SXT	CIP	NIT	NAL	Age	Sex
Nov 25	f02	AMP-SXT	6	6	30	22	29	2	F
Dec 1	f02	AMP-SXT	6	6	31	20	26	4	M
Dec 4	f02	AMP-SXT	6	6	30	21	28	10	M
Dec 6	f02	AMP-SXT	8	6	29	25	27	5	F
Dec 12	f02	AMP-SXT	6	6	30	23	28	2	F
Dec 13	f02	AMP-SXT	6	6	30	21	30	1	M

Signal #4

Date: Feb 6, 2006
 Size: 1 hospital
 Length: 2 days
 Recurrence Interval:
 2.5 years
 Observed Cases: 5
 Expected cases: 0.39
 Relative Risk: 12.8



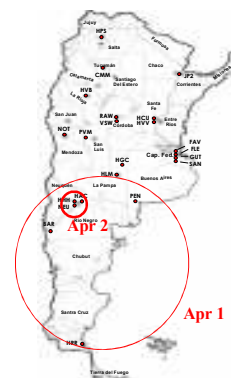
Organisms and Resistance

Date	Spec./clone	Resistance	AMP	SXT	CIP	NIT	NAL	Age	Sex
Feb 5	son-nei	AMP-SXT	6	6	30	22	25	10	F
Feb 5	flex-fl02	none	17	30	30	20	25	2	F
Feb 5	son-nei	AMP-SXT	6	6	30	20	25	2m	M
Feb 6	flex-fl06	AMP-SXT	6	6	28	21	23	6	F
Feb 6	flex-fl06	AMP-SXT	6	6	30	22	26	8	M

Signals #5a,b

Date: April 1, 2007
 Size: 6 hospitals
 Length: 6 days
 Recurrence Interval:
 1.1 years
 Observed Cases: 14
 Expected cases: 4.12
 Relative Risk: 3.4

Date: April 2, 2007
 Size: 3 hospitals
 Length: 7 days
 Recurrence Interval:
 >27 years
 Observed Cases: 14
 Expected cases: 3.53
 Relative Risk: 4.0



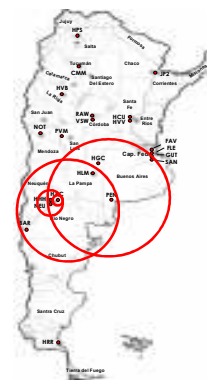
Signal #5c

Date: Apr 17, 2007
 Size: 1 hospital
 Length: 22 days
 Recurrence Interval:
 >27 years
 Observed Cases: 16
 Expected cases: 4.14
 Relative Risk: 3.9



Signal #5d,e,f,etc

Date: Apr & May, 2007
 Size: various
 Length: various
 Recurrence Interval:
 some >27 years
 Observed Cases:
 up to 58
 Relative Risk: various



Conclusions

- The system may have detected some true outbreaks
- A couple of signals are likely chance occurrences, unrelated to any true outbreaks
- The system can only suggest where to look, not whether it is a true outbreak or not
- Adjustments were done for purely spatial and purely temporal variation

Hospital Surveillance

- Brigham and Women's Hospital, Boston
- Years 1997-1999, mimicking a daily prospective surveillance system
- Space-time permutation test statistic
- Organisms or resistance profiles

Hospital Surveillance

“Geography”:

- None: Hospital wide surveillance
- Wards, with neighbors defined both by distance and type of ward (e.g. oncology and bone marrow transplantation wards)
- Service, with neighbors defined by type of service

Hospital Surveillance

Example of a Signal

- Organism: *Candida albicans*
- Date: Nov 10, 2005
- Temporal length: 28 days
- Two wards, Medical Intensive Care Units
- Recurrence interval: > 27 years
- Preceded by signals with lower recurrence intervals

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 physicians, epidemiologists or microbiologists are needed to confirm or dismiss the signals

SaTScan Software

Free. Download from www.satscan.org

Registered users in 116 countries:

1. USA
2. Canada
3. United Kingdom
4. Brazil
5. Italy
- ...
- 100s. Albania, Bhutan, Burma, Fiji, Grenada, Guinea, Iraq, Macao, Madagascar, Malawi, Malta, etc

Acknowledgement

Research funded by:

Alfred P Sloan Foundation
Centers for Disease Control and Prevention
Massachusetts Department of Health
National Cancer Institute
National Institute of Child Health and Development
National Institute of General Medical Sciences:
Modeling Infectious Disease Agent Study (MIDAS)

References

- Kulldorff. A spatial scan statistic. *Communications in Statistics, Theory and Methods*. 26:1481-1496, 1997.
- Hjalmar U, Kulldorff M, Gustafsson G, Nagarwalla N. Childhood leukemia in Sweden: Using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine*, 1996;15:707-715.
- Kulldorff, Heffernan, Hartman, Assunção, Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2(3):e59, 2005.
- Kulldorff and IMS Inc. SaTScan v.7.0: Software for the spatial and space-time scan statistics, 2004. Free: <http://www.satscan.org>