

Statistical methods for Emergency Department chief complaint data:

Temporal scan statistic

To evaluate citywide trends in syndrome visits we use the one-dimensional temporal scan statistic (Wallenstein 1980, Kulldorff 2001). Separate analyses are carried out for each syndrome-age category of interest. The ratio of syndrome visits to 'other' visits (ED visits that do not fall into several infectious disease syndrome categories) during the most recent 3 days is compared to the same ratio during a two-week baseline. The method can scan through temporal clusters of varying sizes up to a pre-set maximum in order to detect sharp one-day spikes and more gradual multi-day increases. We use 3 days so the method examines the previous 1 day, 2 days and 3 days compared to the 14-day baseline. For each cluster a likelihood ratio statistic is calculated that reflects the difference between the observed data and what would be expected under the null hypothesis (no temporal trend). Statistical significance is derived through Monte Carlo hypothesis testing by ranking this likelihood ratio within a distribution of similarly-calculated likelihood ratios from 999 random datasets.

For prospective surveillance the goal is to detect ongoing outbreaks. We therefore restrict the analysis to 'alive clusters only' which include the most recent day of ED visits. To generate a p-value that is adjusted for the multiple comparisons inherent in running daily analyses, we then re-run the program without this restriction and evaluate significance by comparing the likelihood ratio from the 'alive clusters only' analysis to the $p=0.01$ cutoff from this second analysis (see explanation in Kulldorff, 2001).

How we do it

Using SatScan version 2.1.3, a case file is prepared listing the number of syndrome visits citywide on each of the most recent 16 days. A control file lists the number of 'other' visits. A parameter file provides SatScan settings: purely temporal, high-low (detects positive and negative trends), 'alive clusters only'=Yes, maximum temporal size=20% (3 days), Bernoulli model. Finally, a coordinate file is needed, which for this citywide analysis maps all cases and controls to a single citywide coordinate.

[hyperlinks to each file]

CityCases.txt CityControls.txt CityParam.txt CityCoord.txt

SatScan is invoked using the following command which can be typed at the DOS prompt, placed in a batch file, or executed using the SAS 'x' command.

```
c:\SatScanProgramFolder\satscan.exe c:\ParamFolder\Param.txt
```

This analysis evaluates 3 possible temporal clusters: the most recent day compared to the previous 15, the most recent two days compared to the previous 14, and the most recent 3 days compared to the previous 13. For each analysis SatScan determines the observed

and expected cases inside the cluster and during baseline (outside the cluster), and calculates a likelihood ratio statistic.

The program is then re-run a second time with 'alive clusters only'=No in the parameter file. This second analysis evaluates all 3-day clusters throughout the 16-day period for which data are provided, and generates cutoffs that are corrected for the many analyses (multiple comparisons) made during the past 16 days. If the likelihood ratio from the 'alive clusters only' analysis is larger than the 0.01 cutoff from this second 'not alive' analysis, we consider this a statistically significant signal with $p < 0.01$.

Strengths

Software is freely available over internet.

Does not require pre-specifying how an outbreak is distributed over time. Can detect both sharp one-day spikes and more gradual 3-day increases.

P-value is corrected for multiple comparisons that result from examining clusters of different temporal size and, if the 'non-alive' cutoffs are used, from repeating the analysis every day.

Weaknesses

Analysis is univariate in the sense that each syndrome is examined separately. A multivariate model that could detect simultaneous increases in two syndromes (an interaction in time and space) would be a useful enhancement.

Does not correct for multiple comparisons that result from examining several syndrome-age categories each day.

Day-to-day variability in the baseline has no affect. Only the overall ratio of cases to controls during baseline enters into the calculation. This could mask extreme variability in baseline.

The analysis does not make use of historic baseline data beyond the most recent 16 days.

Spatial scan statistic

When an exposure to an infectious agent occurs in a small area of the city, it is reasonable to expect that people living and working in that area are at higher risk for disease. We therefore conduct daily analyses to identify unusual spatial clustering of ED visits for key syndromes. An important assumption underlying this approach is that some people either live or seek care near the source of exposure, and that the inevitable day-to-day mobility of New York City residents and workers will not completely obscure this local signal.

We use an Kulldorff and Mostashari's adaptation of the spatial scan statistic (Kulldorff 1997, Kulldorff 2001) to evaluate clustering in ED visits by hospital address and patient home zip code. We carry out separate analyses for each syndrome-age category of interest. The spatial scan statistic imposes a flexible circular window around each zip code centroid. The circle is allowed to vary in size up to a pre-set maximum (we set this

at 20% of total ED visits). For each unique circle a likelihood ratio statistic is calculated that reflects the difference between the observed and expected syndrome visits inside and outside the circle. A large number of circles are drawn and evaluated, and the cluster with the largest likelihood ratio is the most likely cluster. Secondary clusters that do not overlap with the primary cluster are also identified.

Statistical significance is derived through Monte Carlo hypothesis testing by ranking the cluster likelihood ratio within a distribution of similarly-calculated likelihood ratios based on 999 random datasets. This approach generates p-values that are corrected for the multiple comparisons that result from examining many circles in each analysis (see Kulldorff 1997 and Kulldorff 2000).

How we do it (zip code example)

To run a patient zip code analysis using SatScan version 2.1.3, a case file is prepared containing the observed number of syndrome visits in each zip code on the day of interest (usually the previous calendar day). A 'population' file is also provided and used by the program to calculate the expected number of syndrome visits in each zip code.

Generation of the population file represents our main departure from standard SatScan methodology. The spatial scan statistic was originally developed for the detection of cancer clusters, for which the expected counts can be approximated using the (age-adjusted) census population. This approach cannot be used for ED visits because rates vary widely based on ED utilization patterns, socioeconomic status, access to care and a variety of other factors that may vary geographically and are difficult to measure and control for. In addition, although our system captures 75% of ED visits citywide, coverage is not geographically uniform. We therefore adjust the expected number of visits in each zip code for both purely temporal changes (eg. a citywide increase) and spatial differences seen in the recent 14-day baseline (eg. for a zip code that had a higher proportion of respiratory visits than the rest of the city during baseline the expected will be adjusted up accordingly). To achieve this we provide SatScan with a 'population' for each zip code that is calculated as follows:

1. Baseline proportion = syndrome visits in area during baseline / total visits in area during baseline (*Baseline is the 14-day period ending two days prior to day of analysis*)
2. 'Population' = Baseline proportion * total visits in area on day of analysis * 1000. (*Multiplying by 1000 eliminates decimals while having no affect on spatial distribution*)

A parameter file provides SatScan settings: purely spatial, high clusters only, maximum spatial size=20% (of total ED visits citywide), Poisson model. Finally, a coordinates file provides the longitude and latitude of each zip code centroid.

[hyperlinks to files]

ZipCodeCases.txt ZipCodePop.txt ZipCodeParam.txt ZipCodeCoord.txt

SatScan is invoked using the following command which can be typed at the DOS prompt, placed in a batch file, or executed using the SAS 'x' command (among other methods).

```
c:\SatScanProgramFolder\satscan.exe c:\ParaFolder\Param.txt
```

We consider clusters with $p < 0.01$ significant.

Strengths

Software is freely available over the internet.

Does not require pre-specifying the location or size of the cluster.

P-value is corrected for the multiple comparisons caused by examining a great many circles.

Weaknesses

Although SatScan is capable of flexible scanning in both space and time, we have not yet incorporated a space-time scan analysis into our daily analysis. Instead our analysis considers clustering only in the previous day's data and could miss a cluster that builds slowly over several days.

Our analysis is univariate in the sense that each syndrome is examined separately. A multivariate model that could detect simultaneous increases in two syndromes in the same area would be preferable.

For this spatial-only analysis, the p-value is *not* corrected for the multiple comparisons inherent in running daily analyses, nor is it corrected for the multiple comparisons due to examining several syndrome-age categories.

Cumulative Sums (CUSUM)

CUSUM is a quality control method that has been adapted for aberration-detection in public health surveillance. It has been described elsewhere (Hutwagner 1997). Separate analyses are carried out for each syndrome citywide and for each syndrome by hospital. The unit of analysis is the *proportion* (eg. fever/total) of ED visits—citywide and for each hospital. We have adapted a SAS program developed by Matthew Seaman at the CDC that generates three statistics:

1. **C1:** compares yesterday's proportion to the mean proportion during the previous 7 days. If more than 3 standard deviations above baseline mean a 'C1' signal will occur.

2. **C2:** same as C1 except the 7-day baseline ends 3 days before yesterday. This is the same baseline used in C3.

3. **C3**: for each of the last 3 days, calculates the amount by which the observed proportion exceeds one standard deviation above the baseline mean proportion (7-day baseline ending 3 days before yesterday), effectively ignoring any differences within 1 standard deviation. If the cumulative sum (CUSUM) of these amounts over 3 days is more than 2 standard deviations of the baseline proportions, a 'C3' signal will occur.

Hutwagner LC, Maloney EK, Bean NH et al. (1997) Using laboratory-based surveillance data for prevention: an algorithm for detecting *Salmonella* outbreaks. *Emerg Infect Diseases*, 3(3):395-400.

Kulldorff M. (2001) Prospective time periodic geographical disease surveillance using a scan statistic. *J. R. Statist. Soc., A* 164:61-72.

Kulldorff M. (1997) A spatial scan statistic. *Communs Statist. Theory Meth.*, 26:1481-1496.

Wallenstein S. (1980) A test for detection of clustering over time. *Am J Epidem*, 111:367-372.