Journal of Urban Health: Bulletin of the New York Academy of Medicine Vol. 80, No. 2, Supplement 1 2003

Biosurveillance Applying Scan Statistics with Multiple, Disparate Data Sources

Howard S. Burkom

ABSTRACT *Researchers working on the Department of Defense Global Emerging Infections System (DoD-GEIS) pilot system, the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE), have applied scan statistics for early outbreak detection using both traditional and nontraditional data sources. These sources include medical data indexed by* International Classification of Disease, 9th Revision *(ICD-9) diagnosis codes, as well as less-specific, but potentially timelier, indicators such as records of over-the-counter remedy sales and of school absenteeism. Early efforts employed the Kulldorff scan statistic as implemented in the SaTScan software of the National Cancer Institute. A key obstacle to this application is that the input data streams are typically based on time-varying factors, such as consumer behavior, rather than simply on the populations of the component subregions. We have used both modeling and recent historical data distributions to obtain background spatial distributions. Data analyses have provided guidance on how to condition and model input data to avoid excessive clustering. We have used this methodology in combining data sources for both retrospective studies of known outbreaks and surveillance of high-profile events of concern to local public health authorities. We have integrated the scan statistic capability into a Microsoft Access–based system in which we may include or exclude data sources, vary time windows separately for different data sources, censor data from subsets of individual providers or subregions, adjust the background computation method, and run retrospective or simulated studies.*

KEYWORDS *Biosurveillance, Clustering, Kulldorff, Scan statistics.*

INTRODUCTION

The US Department of Defense Global Emerging Infections System (DoD-GEIS) has developed the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE), which uses syndromic surveillance to create an early warning system for disease outbreaks. ESSENCE monitors more than 100 primary care and emergency clinics in the National Capital Area (NCA) and, since the terrorist attacks in September 2001, has collected data on approximately 100,000 cases per day from US military treatment facilities worldwide. Analysts from DoD-GEIS and The Johns Hopkins University Applied Physics Laboratory have adapted and implemented alerting algorithms for ESSENCE that enable prompt notification of anomalous data counts. ESSENCE II, an extension of the original system, collects civilian and military data in the NCA, with the addition of less-specific, but poten-

Dr. Burkom can be reached at the National Security Technology Department, Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723. (E-mail: howard.burkom@jhuapl.edu)

tially timelier, indicators such as data on over-the-counter (OTC) remedy sales and school absenteeism. Because advances in informatics and communications have permitted substantial increases in the volume and detail of this information, efficient data reduction and interpretation are crucial for rapid recognition of threats to public health.

The principal objectives of ESSENCE II are the early identification, characterization, and tracking of disease outbreaks. ESSENCE II will combine information from widely disparate medical sources, including number of emergency room visits, outpatient visits, and insurance claims, and from nonmedical sources such as counts of OTC remedy sales and school absentees. The time series of daily counts from these sources differ in scale, variance, weekly/seasonal behavior, and other characteristics. Thus, an alerting system that combines data from these sources must tolerate these features and must be able to accommodate data dropouts as well as catch-up reports from individual sources.

We use spatial and temporal data in efforts to improve the promptness of outbreak alerting. A potential problem with sophisticated temporal detectors, however, is choosing the appropriate size and location of the collection region for time series counts. If this region is too small or mislocated, cases may be missed, and the baseline data may not have enough structure, but if the region is too large, the scale and variability of the large-scale time series may reduce sensitivity. We apply spatial-temporal scan statistics in an attempt to localize public health problems promptly. Our early efforts employed the Kulldorff scan statistic[1] as implemented in the SaTScan software[2] of the National Cancer Institute, and research is under way to enhance this capability.

METHODS

SaTScan Concepts

Given a subdivision of the surveillance region into subregions, we wish to find the clusters of subregions with combined data counts that are most unlikely to occur in normal circumstances and to evaluate the significance of these clusters—that is, to estimate how unlikely they are.

Candidate clusters are formed by considering each of a family of circles centered at each of a set of grid points—often taken as the full set of subregion centroids. A candidate cluster is defined as those subregions with centroids that lie in the associated circle. For each grid point, candidate cluster sizes range from a single subregion up to a set of subregions containing a preset maximum fraction of the total case count N. In SaTScan, a statistic is applied to each candidate cluster. Kulldorff's[3] formulation of this statistic is the likelihood ratio *LR*:

$$LR(J) = [\mathrm{O}(J)/E(J)]^{\mathrm{O}(J)} \cdot [(\mathrm{N} - O(J))/(\mathrm{N} - E(J))]^{(\mathrm{N}-\mathrm{O}(J))}$$

where *J* refers to the set of subregions with centroids that lie in a candidate circle, *O*(*J*) is the sum of the observed counts in the subregions included in *J*, *E*(*J*) is the sum of the expected counts in the subregions included in *J*, and N is the total number of cases in the region.

The maximum likelihood cluster is then the set *J** of subregions with the largest LR among the candidate clusters. A *P*-value estimate for the statistical significance of this cluster is determined empirically by ranking the value of *LR*(*J**) among other maximum likelihood ratios, each calculated similarly from a random sample of the N cases based on the expected spatial distribution. Once a set of subregions is associated with a maximal cluster, secondary clusters are chosen from the successively remaining subregions and assigned significance levels using the same ranking.

Adaptations for Processing Multiple Biosurveillance Data Sources

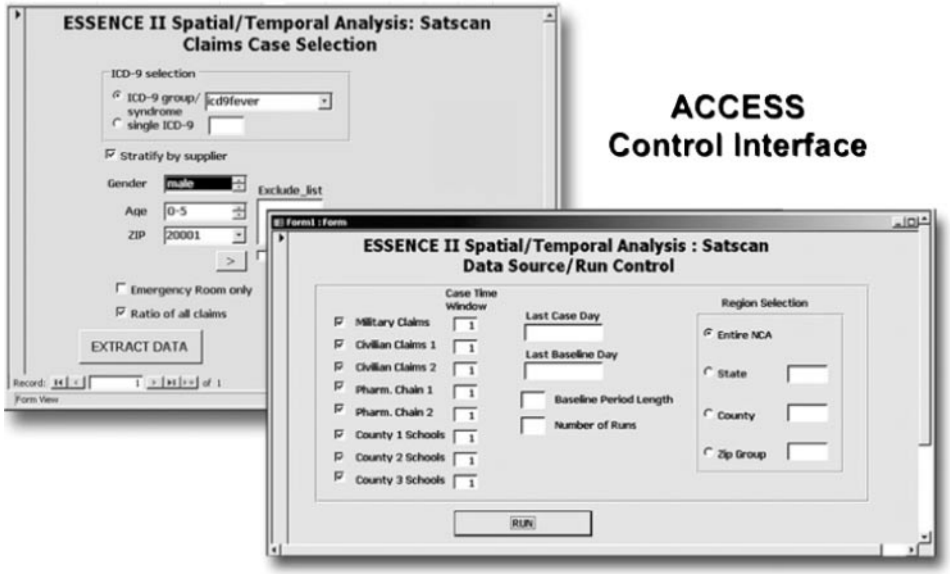
In the conventional use of SaTScan, expected values for the subregions are calculated from the respective populations, assuming uniform spatial incidence; however, counts from most of our data sources are not population based. For example, the distribution of insurance claim data depends on factors such as the distribution of eligible consumers and participating care providers. We derived expected counts both from modeling individual subregion counts and from recent data history. A common technique is to use the spatial distribution of counts from a baseline interval that extends far enough into the past to represent the entire region, yet is recent enough to represent spatial trends.

Our initial approach to combining counts from multiple sources was to treat them as covariates so that we could apply SaTScan directly. This approach requires calculation of expected values for each source in each subregion from source-specific spatial probabilities and case counts. Once expected values are computed, covariate observed and expected counts are summed, and the likelihood ratio statistic is computed. We followed these procedures for multiple sources of medical data treated separately, for absentee counts from different counties (to normalize by county schedule), and for OTC sales from separate store chains. This approach allows us to mix data organized by such variables as patient residence ZIP code, provider location, and store or school address. When adding a new data source, we assign a new covariate number and append the new locations to the aggregate file of spatial coordinates, provided that exact coordinates are not repeated and that each ZIP code or site has a unique identifying string. Expected and observed counts for the new source are then tabulated and included as covariate counts along with counts of the remaining data sources. The spatial clustering includes locations of all the various data sources.

We integrated this multiple-source SaTScan capability into an ACCESS-based system in which we may include or exclude data sources, vary time windows separately for different data sources, censor data from subsets of individual providers or subregions, adjust the background computation method, and run retrospective or simulated studies. Figure 1 shows the control form and the data specification form for outpatient visits.

Data Analysis

Detailed data analysis is necessary before we include a new data source in the surveillance clustering. Without this analysis, applying a scan statistic may produce spurious clusters that can mask the space-time interaction of interest. The general principle is to include the most “signal,” or cases of interest, with the least “noise.” For the recent studies discussed in this article, we used data analysis to improve the expected spatial data distributions in three ways: (1) by the judicious choice of the outcome variable to restrict cases to stably distributed counts, (2) by ignoring counts from subregions with dubious reporting, and (3) by choosing a baseline period that the analysis indicates as both recent and representative. This baseline period may



ESSENCE II Spatial/Temporal Analysis: Satscan
Claims Case Selection

ICD-9 selection
● ICD-9 group/ syndrome: icd9fever
○ single ICD-9
☑ Stratify by supplier
Gender: male
Age: 0-5
ZIP: 20001
Exclude_list
□ Emergency Room only
☑ Ratio of all claims

ESSENCE II Spatial/Temporal Analysis : Satscan
Data Source/Run Control

	Case Time Window
☑ Military Claims	1
☑ Civilian Claims 1	1
☑ Civilian Claims 2	1
☑ Pharm. Chain 1	1
☑ Pharm. Chain 2	1
☑ County 1 Schools	1
☑ County 2 Schools	1
☑ County 3 Schools	1

Last Case Day
Last Baseline Day
Baseline Period Length
Number of Runs

Region Selection
● Entire NCA
○ State
○ County
○ Zip Group

FIGURE 1. Sample output from surveillance combining counts of outpatient visits, over-the-counter antiflu sales, and school absentees.

vary by data type. Other applications may require different methods for spatial estimation, depending on the data sources and surveillance objective.

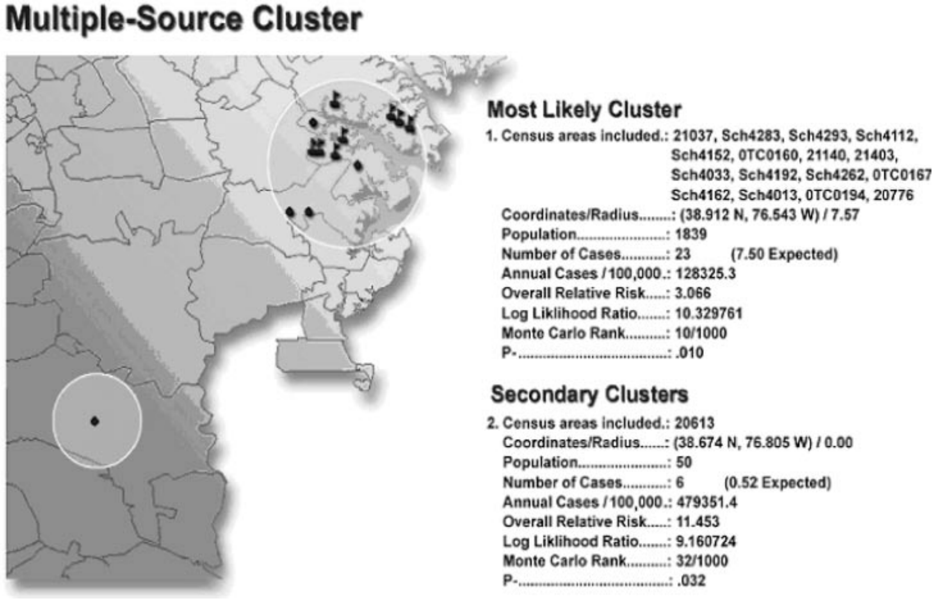
Choice of an outcome variable is important in the use of diagnosis counts for clustering. For medical data, ESSENCE II uses syndromic surveillance, that is, the monitoring of counts of outpatient visits with diagnoses falling in any of seven syndrome groups chosen by DoD-GEIS for surveillance: respiratory, gastrointestinal, fever, dermatologic infectious, dermatologic hemorrhagic, neurologic, and coma. ESSENCE II increments the count for a syndrome group each time a diagnosis code falls in the corresponding list. We examine the spatial and temporal behavior of the various syndrome group counts, especially during cold season, to refine the syndrome groups and subgroups for more sensitive, specific clustering. We also examine each source of data at the local level to reduce noisy temporal behavior that can lead to excessive clustering. For example, we would exclude absentee counts from a school that often skips reporting or that has counts that are especially erratic. For OTC sale data, counts are usually restricted to sales of influenza or diarrhea remedies.

RESULTS

We used this methodology to combine data sources for both retrospective studies of known outbreaks and surveillance of high-profile events of concern to local public health authorities. Figure 2 shows a representative portion of an output file; note that clusters may include sites from any combination of the included data sources.

Simulations for Performance Analysis

In the absence of sufficient authentic disease outbreaks to demonstrate the advantage of using scan statistics with multiple data sources, we used simulations to test



Multiple-Source Cluster

```
Most Likely Cluster
1. Census areas included.: 21037, Sch4283, Sch4293, Sch4112,
                           Sch4152, 0TC0160, 21140, 21403,
                           Sch4033, Sch4192, Sch4262, 0TC0167
                           Sch4162, Sch4013, 0TC0194, 20776
   Coordinates/Radius........: (38.912 N, 76.543 W) / 7.57
   Population.......................: 1839
   Number of Cases...........: 23        (7.50 Expected)
   Annual Cases / 100,000.: 128325.3
   Overall Relative Risk......: 3.066
   Log Liklihood Ratio........: 10.329761
   Monte Carlo Rank...........: 10/1000
   P-...........................................: .010

Secondary Clusters
2. Census areas included.: 20613
   Coordinates/Radius.......: (38.674 N, 76.805 W) / 0.00
   Population.......................: 50
   Number of Cases...........: 6        (0.52 Expected)
   Annual Cases / 100,000.: 479351.4
   Overall Relative Risk......: 11.453
   Log Liklihood Ratio........: 9.160724
   Monte Carlo Rank...........: 32/1000
   P-...........................................: .032
```

FIGURE 2. Control interface for multiple-source surveillance with SaTScan.

the performance of these methods against a variety of modeled outbreaks. We illustrate the methodology for these simulations with a purely spatial Monte Carlo approach. For a particular data source (e.g., counts of outpatient visits with a diagnosis assigned to the respiratory syndrome), assume that we have an expected spatial distribution of these syndrome counts over the subregions (e.g., patient ZIP codes) of the surveillance area. To examine the clusters produced using the scan statistic, we use many iterations of the following procedure:

1. For a set of background cases, compute a spatial case distribution with a multinomial random draw based on expected spatial probabilities.
2. For a test signal, choose an outbreak epicenter in the surveillance region for each test background. Compute a signal probability distribution over the subregions that decays exponentially with the distance from the epicenter. The signal is then a small number of additional cases chosen from this distribution with another multinomial draw.
3. Add the background and signal cases and find the maximal clusters with a spatial scan statistic.
4. Define a true cluster as a computed cluster that contains a subregion with a centroid that is within a threshold distance of the epicenter. Define false clusters as computed clusters that do not satisfy this criterion—clusters away from the epicenter.

For the entire set of clustering runs conducted with this procedure, we then ask: for a threshold value *t*, in what fraction of all runs is there a true cluster with a scan statistic that exceeds *t*, and in what fraction is there a false cluster with a scan statistic that exceeds *t*? By varying this threshold over the values obtained for computed clusters, we obtain a curve, similar to a receiver operating characteristic (ROC) curve, that plots the probability of finding the outbreak versus the probability of a false cluster. (For a discussion of ROC curves, see Ref. 4.)

Figure 3 compares several of these curves. In each case, the number of outbreak cases is 40% of the number of background cases. The thick solid and dashed curves were computed by clustering with respiratory claims and OTC antiflu sales, respectively. The thin dashed curve was computed by clustering with both data sources. The thin solid curve on the logarithmic scale is the 45° line normally used to indicate a completely random detector. For reasonable detection probabilities, we see a substantial gain when the sources are combined.

Dependence on Incidence and Relative Risk of Outbreak Cases

The detector performance shown by the ROC curve depends on the number of outbreak cases added to the data and on the concentration of these cases within the surveillance area. In the above simulation algorithm, the concentration is determined by the decay constant for the exponential spread of the signal cases. In the simulations, we vary the number of added cases and this decay constant to see how the performance of the scan statistic varies with incidence (because these cases are assumed absent from the baseline period) and relative risk in computed clusters. We present an example using the influenzalike illness (ILI) claim counts and antiflu sales in six Monte Carlo simulations with three outbreak sizes, each at two concentration levels. Outbreak sizes in this example are 0.10, 0.15, and 0.20 times the number of background cases. We used two outbreak concentrations. For the more

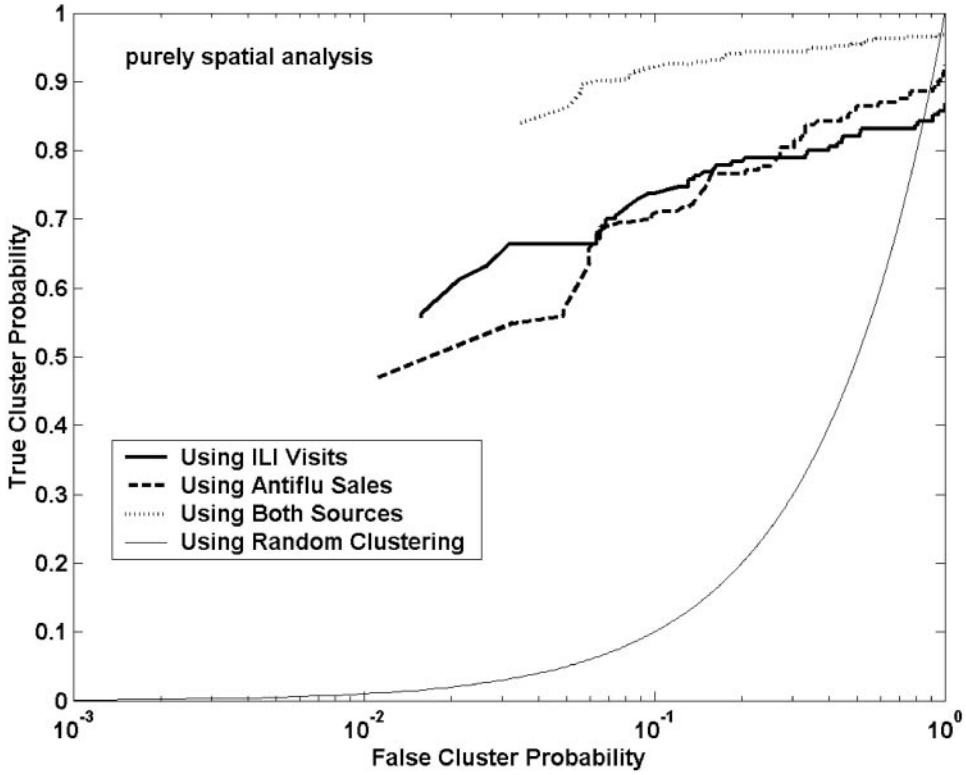


FIGURE 3. Simulation showing potential advantage of spatial clustering using multiple data sources.

diffuse case injections, a circle drawn about the epicenter to include half the injected cases contains about 40 ZIP codes of 190 in the surveillance region, while for the more concentrated injections, this circle contains only about 20 ZIP codes. The ROC curves computed from the six simulations are sampled in the Table, which shows the probabilities of selecting a true cluster for the point on each curve with a 0.05 false cluster probability.

For the diffuse outbreak, the scan statistic with the combination of sources does not show an advantage over using ILI visits alone. For the concentrated outbreak, we see the advantage shown in Fig. 3 for each outbreak size. Similar experiments have suggested that the combination of sources improves the performance of the scan statistic if the injected case concentration falls within limits determined by the outbreak size. If the concentration is too low, none of the methods give high detection probabilities, while if it is too high, they all do well.

This performance analysis technique has several applications. It may be used to assess the marginal surveillance value of a single data source or to investigate how early an outbreak is likely to be detected as the spatial case distribution evolves. We also use it to compare the performance of the likelihood ratio statistic used in SaTScan to other possible scan statistics.

DISCUSSION

This article discusses the application of the method of scan statistics widely used in spatial epidemiology to biosurveillance using multiple, disparate data sources. The success of this application requires an understanding of the spatial and temporal behavior of each source so that we can judiciously choose an outcome variable and calculate the expected spatial distribution of data counts. We implemented this method combining a variety of data sources, and we anticipate increases in early outbreak alerting capability as the number of data sources and promptness of data reporting increase.

Limitations and Caveats

Clusters identified by SaTScan or by our derived methods should be understood as approximate locations of concentrated data counts that may indicate an outbreak of disease. The statistical significance and persistence of these clusters should be used to evaluate their importance. They are also valuable as cues for and corrobora-

TABLE. Empirical probabilities of finding clusters of injected cases gven a .05 probability of computing a false cluster, with varying outbreak size and concentration

Injected risk type	Ratio of injected to background cases	Influenzalike illness visits	Over-the-counter sales	Visits and sales combined
Diffuse	.10	.31	.23	.31
	.15	.15	.17	.25
	.20	.51	.37	.50
Concentrated	.10	.35	.28	.48
	.15	.39	.46	.70
	.20	.66	.58	.86

tion of other surveillance measures. Several potential problems for the data fusion approach presented above must also be understood:

1. Because the scan statistic numerators are formed by adding counts of disparate sources, a data source with highly variable or relatively large counts may mask signals in sources with smaller or more stable counts.
2. The addition of sources causes the computation of more clusters with just a few cases where nearly none are expected. A strategy for when to abandon such clusters as "epidemiologically insignificant" is necessary. Unless treated as exceptions, these small clusters raise the false cluster probabilities on the ROC curves.
3. For a realistic simulation, data analysis is needed to estimate the relative effects of an outbreak on different data sources and the expected time delay for the effect to appear in each source.

A Stratified Scan Statistic

We implemented a simple modification to the data fusion approach to treat the problem of different data scales or variances noted in the first caveat above. Instead of computing $\log(LR)$ using numerators and denominators summed over the sources, we compute $\log(LR_j)$ separately for each source j and treat the sum of these logarithms as the scan statistic. The downloadable SaTScan software does not allow this modification, but our preliminary experience with it has been encouraging. Summing the log likelihood values does lessen the problem of mismatched data scales; the intuitive drawback is that computing separate LR_j values decreases the power to detect a faint outbreak with slight increases in multiple sources. However, ROC comparisons of this stratified method with the covariate method suggest at most a slight loss in the power to detect combined clusters.

Another potential advantage of taking individual $\log(LR_j)$ values is that the statistic may be modified by weighting these values; such weighting cannot be done in the covariate method because the SaTScan numerators must be original data counts. Furthermore, the strategy of summing logarithms for separate data sources may be extended to other candidate scan statistics, as in Ref. 5.

We are continuing efforts to improve the use of scan statistics for multiple data sources through data analysis, mathematical fusion methods, and alternate scan statistics.

ACKNOWLEDGEMENT

This research is sponsored by the Defense Advanced Research Projects Agency and managed under Naval Sea Systems Command contract N0024-98-D-8124.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, NAVSEA, or the US government.

REFERENCES

1. Kulldorff M. Spatial scan statistics: models, calculations, and applications. In: Glaz J, Balakrishnan N, eds. *Scan Statistics and Applications*. Boston, MA: Birkhauser; 1999: 303–322.

2. Kulldorff M. *SaTScan* [computer program]. Version 2. Available at: http://srab.cancer.gov/satscan/. Accessed on: September 1, 2002.
3. Kulldorff M. A spatial scan statistic. *Commun Stat Theory Meth*. 1999;26:1481–1496.
4. Swets JA, Pickett RM. *Evaluation of Diagnostic Systems: Methods From Signal Detection Theory*. New York, NY: Academic Press; 1982.
5. Edgington ES. A normal curve method for combining probability values from independent experiments. *J Psychol*. 1972;82:85–89.