

Multivariate Scan Statistics for Disease Surveillance

Martin Kulldorff* Farzad Mostashari Luiz Duczmal
Katherine Yih Ken Kleinman Richard Platt

Abstract

In disease surveillance, there are often many different data sets or data groupings for which we wish to do surveillance. If each data set is analyzed separately rather than combined, the statistical power to detect an outbreak that is present in all data sets may suffer due to low numbers in each. On the other hand, if the data sets are added by taking the sum of the counts, then a signal that is primarily present in one data set may be hidden due to random noise in the other data sets.

In this paper, we present an extension of the spatial and space-time scan statistic that simultaneously incorporates multiple data sets into a single likelihood function, so that a signal is generated whether it occurs in only one or in multiple data sets. This is done by defining the combined log likelihood as the sum of the individual log likelihoods for those data sets for which the observed case count is more than the expected.

Using data from the National Bioterrorism Syndromic Surveillance Demonstration Project, we illustrate the new method using physician telephone calls, regular physician visits and urgent care visits by Harvard Pilgrim Health Care members cared for by Harvard Vanguard Medical Associates, a large multi-specialty group practice in Massachusetts. For upper and lower gastrointestinal illness, there were on

*Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, 6th Floor, Boston, MA 02215, USA. Email: martin_kulldorff@hms.harvard.edu, Tel: 617-509-9757, Fax: 617-859-8112.

average 20 telephone calls, 9 urgent care visits and 22 regular physician visits per day. The strongest signal was generated by a single data set and due to a familial outbreak of pinworm disease. The second and third strongest signals were generated by the combined strength of two of the three data sets.

Keyword: spatial statistics, syndromic surveillance, clusters, disease outbreaks.

1 Introduction

The spatial and space-time scan statistics [1, 2, 3, 4] have become popular methods in disease surveillance for the detection of disease clusters. The standard approach is to look at a single disease or health outcome, such as leukemia incidence [5], breast cancer mortality [6] or late stage prostate cancer [7]. Very often though, it is of interest to simultaneously look at multiple outcomes. For example, if we are interested in investigating whether there are any geographical clusters of childhood leukemia, it is not easy to know a priori whether to look for clusters of acute lymphocytic leukemia (ALL), of acute myelogenous leukemia (AML), of chronic leukemia or of all leukemia combined. In surveillance, as opposed to the evaluation of specific epidemiological hypotheses, it is then natural to simultaneously evaluate the geographical distribution of any combination of these outcomes.

One very interesting application of the space-time scan statistic is in syndromic surveillance for the early detection of disease outbreaks [8, 9]. In syndromic surveillance, symptoms such as fever or vomiting are used for the early detection of disease outbreaks. The primary reason for using symptoms rather than diagnosed diseases is that it can take many tests and several days to establish a firm diagnosis. It is hence thought that a surveillance system based on symptoms such as diarrhea or vomiting may allow public health officials to detect a disease outbreak at the earliest time possible. There are three major reasons for taking a multivariate approach to syndromic surveillance:

1. No single data source captures all the individuals in the outbreak. Depending on the disease, some will go to their pharmacy and buy an over-the-counter medication, some will call their physician or a nurses hot-line, while other may visit their regular physician, go to a hospital emergency room or call the ambulance.

2. Some diseases will typically manifest itself with a single symptom like diarrhea, while other disease can cause a wide variety of different symptoms in different individuals.
3. Some diseases such as rotavirus affect mainly children, while other diseases affects the population at large or mainly the elderly. If it is a work-site exposure causing the outbreak, it will be mainly working-age adults that are affected.

Since we do not know what disease we are looking for, we do not know in which of these groupings that we should look for an outbreak: which combination of data sources, symptoms and/or age groups?

One simple approach is to sum up the counts from in all the data sets and do the analysis on the combined data. The drawback is that if an outbreak is only present in one of the data sets it may be hidden by the random variation present in the other data sets, with an absent or delayed signal as a result. Another approach to dealing with multiple data sets is to analyze each separately, and then possibly use a Bonferroni type adjustment for the multiple testing. The drawback of this approach is the potential lack of power and consequently a delay in the signal if the outbreak is present simultaneously in more than one of the data sets. Hence, what we want to do is to simultaneously evaluate both single data sets and various combinations of multiple data sets.

Even if a signal is present in multiple data sets, there is often different informational content in observations from different data sets. For example, a disease outbreak may generate both additional pharmacy sales and additional ambulatory care visits, but ten more diarrhea medication sales in addition to a daily count of 1000 is not as informative as ten more diarrhea related emergency department visits in addition to an average daily count of 20. By simply adding the counts, the difference in informational content is lost.

The idea of simultaneous analysis of multiple data sets or groupings for disease surveillance is not new. In the purely temporal setting, Shmueli and Fienberg [10] discuss a number of options. Dillard and Shmueli have applied wavelets to multiple time series [11]. Naus and Wartenberg [12] have developed purely temporal scan statistics for two data types, but their purpose is different, to find clusters with a minimum number of both types of events. To deal with multiple groupings, Wong et al. [13] proposed the WSARE system: *‘What’s Strange About Recent Events?’*, which searches for outbreaks in various groupings of for example age and gender. In this

system, geographical areas such as county or census tract may be one of the groupings, but the spatial locations of those areas are ignored. Kulldorff et al. [14] developed a tree-based scan statistic to do surveillance on groupings that can be pre-classified into a hierarchical tree structure.

In a syndromic surveillance paper on applying the space-time scan statistic with multiple disparate data sources, Burkom [15] took the approach of adding the counts as well as the expected counts, but he also mentions the possibility of instead adding the log likelihoods. That is a nice way to handle the different informational content in different data sets, but great care needs to be taken when doing the summation. Rather than doing a raw summation over all data sets, separate summations should be done for the data sets with more versus less cases than expected in the scanning window. In this paper, we propose a multivariate scan statistic with the ability to detect clusters in either one or in a combination of data sets, without having to pre-specify which ones. We also propose a way to adjust a univariate scan statistic taking the differential informational content in different data sets into account.

2 The Univariate Scan Statistic

We first briefly review the existing theory of univariate scan statistics with variable size window [16, 2], starting with the space-time version. Details are provided in the cited publications. Let c_{st} be the observed number of cases in the geographical location s during time period t . Let n_{st} be either the population or the expected number of cases in location s during time period t . Let $C = \sum_{s,t} c_{st}$ be the total number of cases and let $N = \sum_{s,t} n_{st}$ be the total population/expected cases. The expected may be calculated in different ways adjusting for various covariates such as age, gender, urbanicity, day-of-week or seasonal effects [2, 17, 18].

The space-time scan statistic is defined through a huge number of overlapping cylinders [19]. For each cylinder z , a log likelihood ratio $LLR(z)$ is calculated, and the test statistic is defined as the maximum LLR over all cylinders. The collection of cylinders will depend on the application. The circular base defines a geographical area with the radius varying continuously from zero up to some upper limit that is often defined so that the circle contains at most 50 percent of the population at risk. It can also be defined as for example 10 kilometers. The height of the cylinder reflects a shorter or

longer time period, and could be anything from a single day to several years. For each choice of circle all choices of the temporal height is considered and vice-versa, so that the scanning is done over short and fat cylinders in the shape of a plate, tall and thin cylinders in the shape of a pen, and everything in between. Let $c_z = \sum_{s,t \in z} c_{st}$ be the number of cases in cylinder z and let $n_z = \sum_{s,t \in z} n_{st}$ be the population or the expected number of cases in cylinder z as applicable.

The mathematical formulation of the log likelihood ratio depends on the probability model used. For a Poisson model, where the counts are generated from a Poisson distribution, the log likelihood ratio is

$$LLR(z) = \left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{C - c_z}{C - n_z}\right)^{C - c_z}$$

where n_z represents expected counts and $C = N$ [2]. The definition of the scan statistic is

$$T = \max_z LLR(z).$$

Other probability models have also been used for scan statistics. A Bernoulli model is used for 0/1 case-control type data [2], an ordinal model for ordinal data [20], an exponential model for survival data [21] and a space-time permutation model for looking at space-time interaction clusters when only case data is available [9]. The latter also uses the Poisson distribution but it is an approximate rather than an exact log likelihood ratio. For the purpose of this paper, it does not matter which probability model we work with. Only the $LLR(z)$ function is different, while the principles for dealing with multiple data sets is the same.

Once the test statistic has been calculated, the next step is to evaluate its statistical significance. This is done by generating a large number (typically 999) of random data sets generated under the null hypothesis of no clustering, and then calculating the value of the test statistic for each of those data sets as well. The most likely cluster for the real data set is then significant at the 0.05 alpha level if the value of its scan statistic is among the 5 percent highest among the one real and many random data sets [22]. More generally, its Monte Carlo hypothesis testing p-value is defined as $p = r/(1 + sim)$ where r is the rank and sim is the number of simulated Monte Carlo replications generated under the null hypothesis. In this way, the p-value of the scan statistic is adjusted for the multiple testing inherent in the many cylinder locations and sizes evaluated.

When there is only one time period, the space-time scan statistic simplifies to the purely spatial scan statistic. Likewise, if there is only one spatial location, then it simplifies to the purely temporal scan statistic. Hence, while we will describe the subsequent theory in the framework of the space-time scan statistic, everything applies to the purely spatial and purely temporal scan statistics as well. Other shapes of the scanning window than circles and cylinders are also possible [2, 23, 24, 25], but again, the principles for dealing with multiple data sets is the same.

3 The Multivariate Scan Statistic

Sometimes it is interesting to simultaneously search for and evaluate clusters in more than one data set. The multivariate scan statistic with multiple data sets solves this problem, as follows.

1. For each cylinder, the log likelihood ratios are calculated for each data set, and it is noted whether the observed number of cases is larger or smaller than expected.
2. For each cylinder, the log likelihood ratios for the data sets with more than expected number of cases is summed up. This sum is one of two likelihoods for that particular cylinder. The second likelihood is the sum of all the log likelihood ratios for the data sets with fewer than expected cases.
3. The maximum of all the summed log likelihood ratios, taken over all the cylinders, constitutes the most likely cluster. This maximum is the definition of the multivariate scan statistic.

We will now write this down in mathematical notation. For the Poisson model, let

$$LLR_i(\text{high}, z) = \left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{C - c_z}{C - n_z}\right)^{C - c_z} I(c_z > n_z)$$

and

$$LLR_i(\text{low}, z) = \left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{C - c_z}{C - n_z}\right)^{C - c_z} I(c_z < n_z)$$

be the log likelihood ratio for high and low clusters respectively for cylinder z in data set i . Note that for a specific cylinder and data set, only one of these can be non-zero. The test statistic can now be written as

$$T = \max_z \max \left(\sum_i LLR_i(\text{high}, z), \sum_i LLR_i(\text{low}, z) \right)$$

Note that when defined in this way, the most likely cluster may either be a 'high cluster' with more cases than expected or a 'low cluster' with less cases than expected. If there is only interest in one of these two, only one of the sums in the second step above is used. That is, when search only for high clusters,

$$T = \max_z \sum_i LLR_i(\text{high}, z).$$

Note that with the multivariate scan statistic, we do not only adjust for the multiple testing inherent in the many cylinder location and sizes evaluated in the same way as explained for the univariate case, but we also automatically adjust for the multiple testing inherent in the various combinations of data sets that can generate the most likely cluster. This is done by taking the maximum over all data set combinations when taking finding the maximum log likelihood in the random data sets in exactly the same way as for the real data.

4 Adjusting for Different Informational Content in Different Data Sets

When data is classified by categorical covariates such as different age groups, disease classifications or data sources, the simplest approach is to sum up both the cases and the population/expected counts, and then use a univariate scan statistic. Often though, it is important to adjust the analysis for these covariates. For example, an unadjusted spatial scan statistic analysis of cancer mortality in the United States would reveal significant clusters in Florida, since many old people live there and older people are at much higher risk of cancer. The standard way to adjust the Poisson based scan statistic for covariates is to calculate the expected counts through indirect standardization. Let c_{ist} and pop_{ist} be the number of cases and the population in covariate group i in location s in time period t respectively. The

covariate adjusted expected count is then $n_{st} = \sum_i pop_{ist} C_i / POP_i$, where C_i is the total number of cases in category i and POP_i is the total population in category i . With this type of adjustment, there is no differentiation in the informational content between cases in different covariate groups. That is, a fifty year old leukemia patient in a location provides no more and no less evidence for a cluster there than a eighty year old patient. This equality is often appropriate but not always. For example, in influenza mortality surveillance, one additional death of a thirty year old person would provide more evidence of an outbreak with serious public health consequences than one additional death of an eighty year old person. This is because influenza mortality is common among the elderly during a normal flu season, while many influenza deaths among the younger population may signal the arrival of a very serious strain, so that there is more 'informational content' in the death of the thirty year old person.

To adjust for different informational content in different covariate groups, we need to adjust the analysis in a different manner, as follows.

1. Divide the data into different data sets, one for each covariate group.
2. For each cylinder z calculate the log likelihood ratio for each data set, and note whether the number of cases is higher or lower than expected.
3. For each cylinder, the log likelihood ratios for the data sets with more than expected number of cases is summed up. Likewise, sum up all the log likelihood ratios for the data sets with fewer cases than expected.
4. For each cylinder, subtract the second sum from the first, to create the log likelihood for high clusters with an excess number of cases. Subtract the first sum from the second, to create the log likelihood for low clusters with fewer cases than expected. The larger of these is the log likelihood for this particular cylinder.
5. The maximum of all the log likelihoods, taken over all the cylinders, constitutes the most likely cluster. This maximum is the definition of the informational content adjusted scan statistic.

When there is only interest in clusters with high rates, the same procedure is used, except that only the first subtraction in step four is performed. In mathematical notation, the scan statistic is then for the Poisson model

expressed as

$$T = \sum_i LLR_i(high, z) - \sum_i LLR_i(low, z)$$

where $LLR_i(high, z)$ and $LLR_i(low, z)$ are defined as above. When searching for both high and low clusters

$$T = \max \left(\sum_i LLR_i(high, z) - \sum_i LLR_i(low, z), \sum_i LLR_i(low, z) - \sum_i LLR_i(high, z) \right)$$

Note that the two types of adjustment may be used simultaneously for different covariates. For example, in the influenza example, we could use the indirect standardization approach to adjust for gender and the multiple data set approach to adjust for age groups. We would then treat each age group as a separate data set, and for each data set we would calculate the gender adjusted expected counts for each location and time period combination.

5 Example: Syndromic Surveillance in Boston

5.1 Patient Encounter Data

Using data from the National Bioterrorism Syndromic Surveillance Demonstration Project, we illustrate the multivariate scan statistic using physician telephone calls, regular physician visits and urgent care visits by the Harvard Pilgrim Health Care health plan members cared for by Harvard Vanguard Medical Associates, a large multi-specialty group practice in Massachusetts. Data on patient encounters (visits or calls), including demographic information and diagnostic codes, are recorded electronically as part of routine patient care, usually on the same day. Each night, encounters with codes of interest are extracted automatically from clinical data systems. The extracted encounter files are created to uniform specifications and are kept on a directory accessible to software ("the console") provided by the data center.

The console maps encounters to syndromes, in our case upper gastrointestinal (GI) illness, and then identifies illness episodes by omitting individuals' encounters in any syndrome that occur within 42 days of an earlier visit in the same syndrome. Episodes are mapped to the individuals' zip code of residence, and a single file is created containing counts of new episodes in each syndrome and zip code for each day. In addition, historical episode files are created, which provide a basis for modeling end evaluation work of the

type presented in this paper. During the processing of encounter files into episode files, the console produces encounter lists, containing demographic and clinical information, which remain at the originating site, where they are available in the event of a query from public health authorities.

For this paper, we use historical data from 2002, mimicing a real-time prospective surveillance system with daily analyses. On average, there were about 20 new upper GI episodes per day through telephone call, about 22 regular physician visits and about 9 urgent care visits.

5.2 Choice of Scan Statistic and Parameter Settings

For this example, the prospective space-time permutation scan statistic [9] was used for the early detection of disease outbreaks. This scan statistic uses only case data and the expected values for each location and time period is calculated by conditioning on the marginals. For example, if 1000 out of a total of 10,000 cases (=10%) are in location A and there is a total of 100 cases during day d , then the expected number of cases in location A on day d is ten percent of 100, equal to 10. If there is a total of 200 cases on the next day, then the expected is 20 in location A on that day. Based on these expected counts, the space-time permutation scan statistic uses the Poisson based likelihood described above to define the test statistic.

To define the collection of cylinders, we used 3 days as the maximum temporal window size. That is, the height of the cylinder is either 1, 2 or 3 days. Since we are doing prospective surveillance, we are only interested in ongoing outbreaks, so the last day of observation is always included as one of the days in the cylinder. The circular base of the cylinder was defined so that its centroid coincided with one of the zip-code centroids. For each centroid, we evaluated all circles with a radius between zero and a centroid dependent maximum so that at most 25 percent of all cases in the data set was included in the circle.

A space-time scan statistic analysis was performed for each of the 122 days from September 1 to December 31, 2002, using data from January 1, 2002 until and including the 'day of the analysis'. Calculations were performed using the freely available SaTScan software [26].

5.3 Results

The results of the multivariate prospective space-time permutation scan statistic are presented in Table 1. The strongest signal was on October 18, with the cluster consisting of a single day and a single zip-code. On that day in that zip-code, there were five individuals with upper GI illness compared to 0.04 expected. With a recurrence interval (RI) greater than 1000 days ($p = 0.001$), we would by chance only expect to see an outbreak of this magnitude about once every three years. Hence, this cluster is unlikely to be a chance occurrence.

A detailed investigation of the cluster revealed that all five were diagnosed with pinworm disease, one of the more unusual forms of GI illness. The investigation also revealed that all five were members of the same family, with three of them being children. Such a signal is clearly not of public health importance, but it does show the ability of the system to pick up very small clusters if they are highly concentrated in time and space.

A set of three consecutive signals occurred on December 20 ($p = 0.002$, RI=500 days), December 21 ($p = 0.001$, RI=1000 days) and December 22 ($p = 0.002$, RI=500 days). These three signals were generated by the same cluster, the first two containing the same seven zip-code areas during two and three days respectively, and the third containing three of those seven zip-code areas. The next strongest signal occurred on October 26 ($p = 0.007$, RI=142 days). The individuals in both the October and December clusters had a fairly wide variety of different gastrointestinal illness, and it is not clear whether or not there were a common etiological factor for the cases in these two clusters.

To compare with the multivariate scan statistic, we also ran separate analyses for each of the three data sets, again for each of the 122 days. The October 18 analysis of the telephone data also picked up the 5 person pinworm cluster ($p = 0.001$, RI=1000), while obviously, the regular visit and urgent care data sets did not. The December 20 cluster was picked up by both the telephone calls ($p = 0.03$, RI=32 days) and the regular physician visit ($p = 0.003$, RI=333 days) data sets, also on December 20, with similar size signals on the subsequent two days. The October 26 cluster was not picked up by any of the separate data sets ($p > 0.05$, RI < 14 days).

Date	Signal Characteristics		Telephone Calls			Regular Visits			Urgent Care			p=	RI
	#zips	#days	obs	exp	RR	obs	exp	RR	obs	exp	RR		
Oct 18	1	1	5	0.02	250	0	0.02	0	0	0	n/a	0.001	1000
Dec 20	7	2	7	1.24	5.6	7	1.18	5.9	2	1.12	1.8	0.002	500
Dec 21	7	3	9	1.64	5.5	7	1.20	5.8	2	1.60	1.3	0.001	1000
Dec 22	3	3	5	0.68	7.4	4	0.23	17.3	0	0.98	0	0.002	500
Oct 26	5	2	5	0.50	10.0	3	0.40	7.5	0	0.39	0	0.007	142

Table 1: The strongest signals of surveillance for upper gastrointestinal illness in Boston, during September 1 to December 31, 2002, as generated by the multivariate space-time permutation scan statistic. All signals with a recurrence intervals (RI) greater than 120 days are listed. By chance we would expect approximately one signal during this four month period. The signals on December 20-22 are for the same outbreak. RR = relative risk.

6 Discussion

We have presented a multivariate scan statistic and illustrated its use for syndromic surveillance and the early detection of disease outbreaks. The multivariate approach can also be used for purely temporal and purely spatial surveillance in a wide variety of settings. For example, in geographical cancer surveillance, one may treat acute lymphoblastic leukemia as one data set, acute myelogenous leukemia as another and chronic leukemia as a third. The multivariate scan statistic is then able to detect clusters due to only one of the three types or two or three of them combined, adjusting for the additional multiple testing due to the multiple combinations of the three data sets.

When using the multivariate scan statistic, one automatically adjusts for the differential informational content when combining different groupings. Hence, we are really modifying the univariate scan statistic in two different ways in this paper. It is also possible to adjust for the informational content without using a multivariate approach, as described in Section 4. The converse is also possible. It is possible to construct a multivariate scan statistic without the informational content adjustment. To do this, one would for each cylinder add the cases and the population/expected separately and then calculate the log likelihood based on these sums rather than summing the log likelihoods. We have not tried this approach.

The ability to adjust scan statistics for different informational content is

very interesting in itself. Of particular interest in this regard is the ability to adjust the Bernoulli based scan statistics for covariates, as that has not been possible before.

The methods described in this paper are computer intensive, both because of the nature of the scanning window and because of the need to evaluate the test statistic for 999 or more random replicas of the data set. To make it easy to apply, the methods have been incorporated into the freely available SaTScan software [26]. Using this software, it is possible to run purely temporal, purely spatial as well as space-time multivariate scan statistics, using either a Poisson, Bernoulli or space-time permutation probability model. It is also possible to do the informational content adjustment for all these scenarios.

Acknowledgment

This work was funded by the Alfred P. Sloan Foundation (MK,FM,LD) and the National Center for Infectious Diseases, Centers for Disease Control and Prevention, grant number CCU115079 (MK,KY,KK,RP). Their support is greatly appreciated.

References

- [1] J Naus. Clustering of random points in two dimensions. *Biometrika*, 52:263–267, 1965.
- [2] M Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26:1481–1496, 1997.
- [3] J Glaz and N Balakrishnan, editors. *Scan Statistics and Applications*. Birkäuser, 1999.
- [4] J Glaz, J Naus, and S Wallenstein. *Scan Statistics*. Springer, 2001.
- [5] U Hjalmars, M Kulldorff, G Gustafsson, and N Nagarwalla. Childhood leukemia in sweden: Using gis and a spatial scan statistic for cluster detection. *Statistics in Medicine*, 15:707–715, 1996.

- [6] CE Hsu, HE Jacobson, and F Soto Mas. Evaluating the disparity of female breast cancer mortality among racial groups - a spatiotemporal analysis. *International Journal of Health Geographics*, 3:4, 2004.
- [7] A Klassen, M Kulldorff, and F Curriero. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *International Journal of Health Geographics*, 4:1, 2005.
- [8] R Heffernan, F Mostashari, D Das, A Karpati, M Kulldorff, and D Weiss. Syndromic surveillance in public health practice: The new york city emergency department system. *Emerging Infectious Diseases*, 10:858–864, 2004.
- [9] M Kulldorff, R Heffernan, J Hartman, R Assunção, and F Mostashari. A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine*, 2:e59, 2005.
- [10] G Shmueli and SE Fienberg. Current and potential statistical methods for monitoring multiple data streams for bio-surveillance. In A Wilson and D Olwell, editors, *Statistical Methods in Counter-Terrorism*. 2004.
- [11] B Dillard and G Shmueli. Simultaneous analysis of multiple time series using two-dimensional wavelets. *Manuscript*, 1:1, 2004.
- [12] JI Naus and D Wartenberg. A double-scan statistic for clusters of two types of events. *Journal of the American Statistical Association*, 92:1105–1113, 1997.
- [13] WK Wong WK, A Moore, G Cooper, and M Wagner. Wsare: What’s strange about recent events? *Journal of Urban Health*, 80:i66–i75, 2003.
- [14] M Kulldorff, Z Fang, and S Walsh. A tree-based scan statistic for data-base disease surveillance. *Biometrics*, 9:641–646, 2003.
- [15] H Burkom. Biosurveillance applying scan statistics with multiple disparate data sources. *Journal of Urban Health*, 80:–57–i65, 2003.
- [16] CR Loader. Large-deviation approximation to the distribution of scan statistics. *Annals of Applied Probability*, 23:751–771, 1991.

- [17] M Kulldorff, E Feuer, B Miller, and L Freedman. Breast cancer in northeastern united states: A geographic analysis. *American Journal of Epidemiology*, 146:161–170, 1997.
- [18] K Kleinman, A Abrams, M Kulldorff, and R Platt. A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, page in press, 2005.
- [19] Kulldorff M, Athas W, Feuer E, Miller B, and Key C. Evaluating cluster alarms: A space-time scan statistics and brain cancer in los alamos. *American Journal of Public Health*, 88:1377–1380, 1998.
- [20] I Jung, M Kulldorff, and A Klassen. A spatial scan statistic for ordinal data. *Manuscript*, 1:1, 2005.
- [21] L Huang L, M Kulldorff, and D Gregorio. A spatial scan statistic for survival data. *biometrics*,. *Manuscript*, 1:1, 2005.
- [22] M Dwass. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28:181–187, 1957.
- [23] GP Patil and C Taillie. Geographic and network surveillance via scan statistics for critical area detection. *SS*, 18:457–465, 2003.
- [24] L Duczmal and R Assuncao. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis*, 45:269–286, 2004.
- [25] T Tango and K Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4:11, 2005.
- [26] M Kulldorff and Information Management Services Inc. *SaTScan v5.1: Software for the spatial and space-time scan statistics*. www.satscan.org, 2004.