Statistics in Medicine

Research Article

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A spatial scan statistic for compound Poisson data

Rhonda J. Rosychuk and Hsing-Ming Chang[*†]

The topic of spatial cluster detection gained attention in statistics during the late 1980s and early 1990s. Effort has been devoted to the development of methods for detecting spatial clustering of cases and events in the biological sciences, astronomy and epidemiology. More recently, research has examined detecting clusters of correlated count data associated with health conditions of individuals. Such a method allows researchers to examine spatial relationships of disease-related events rather than just incident or prevalent cases. We introduce a spatial scan test that identifies clusters of events in a study region. Because an individual case may have multiple (repeated) events, we base the test on a compound Poisson model. We illustrate our method for cluster detection on emergency department visits, where individuals may make multiple disease-related visits. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: cluster detection; spatial scan; compound Poisson; surveillance

1. Introduction

Spatial cluster detection (SCD) methods provide tools to find proximities where certain events occur significantly more (or less) often than expected. SCD methods are popular in surveillance of diseases (e.g. [1–6]), studying growth pattern of vegetation over landscapes (e.g. [7, 8]), and in crime pattern analysis (e.g. [9]). The spatial pattern of disease spread or other health outcomes often is of interest to health authorities. Authorities typically collect administrative health data that can be used to study the epidemiology of diseases, patterns of health services provided, and the relationships between potential sources and outcomes of diseases. Databases housing such health data can also be utilized for evaluation of the efficiency and effectiveness of public health services. SCD methods can be employed with health data as surveillance tools to help monitor, in an objective and statistically sound manner, health outcomes across a geographic region.

Several authors have contributed to the development of and investigated the properties of SCD methods (see [10–13] for reviews). General tests are designed to detect clusters within the overall pattern of disease in a complete region [1]. Under

Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada
*[*] *Correspondence to: Hsing-Ming Chang (Postdoctoral Fellow), Department of Pediatrics, University of Alberta, 3-077B Edmonton Clinic Health Academy, 11405 87 Avenue NW, Edmonton, Alberta, Canada T6G 1C9.*
[†] *E-mail: hsingmin@ualberta.ca*

Contract/grant sponsor: Canadian Institutes of Health Research

these tests, the cases of disease are assumed to occur at random: each individual in the population has an equal chance of developing the disease. For these tests, no specific alternative distribution for the cases is hypothesized.

Many of the available SCD methods assume that the geographic areas have comparable population sizes and are not applicable to data consisting of areas with diverse population sizes. Our applications involve diverse population sizes and we focus our review on a few key methods. Often methods can be summarized by either looking at areas of constant population and comparing the number of cases [2, 14, 15] or looking for a specified number of cases and comparing the underlying population sizes [1]. Turnbull et al. [15] take the former approach by combining geographic areas into circles that constitute constant numbers of individuals at risk, calculating the number of cases in each circle, and assessing statistical significance through Monte Carlo simulations. Kulldorff and Nagarwalla [2] generalize this approach with a likelihood ratio test that assesses if the individuals in a zone (circle) have greater disease risk than those outside the zone. By maximizing the likelihood function over connected subgraphs of the study region, a similar likelihood ratio test is provided by Duczmal and Assunção [16]. Methods by Besag and Newell [1] and Tango [17] identify areas with a tendency to cluster. Besag and Newell [1] combine regions with nearest neighbors and compare the number of neighbors that must be combined to contain a pre-specified number of cases. A chi-square statistic based on the discrepancy between observed and expected relative frequencies and a "closeness" measure is proposed by Tango [17].

The SCD methods described are all based on detecting excess cases of disease and more recent developments and extensions have included the detection of excess events related to a particular condition or disease (e.g., [4, 6]). These tests are based on a strategy similar to Besag and Newell's [1] general test. We now consider a spatial scan for compound Poisson data to detect geographic areas with excess events. The test is in the same spirit as the spatial scan test introduced by Kulldorff and Nagarwalla [2]. Our method uses a compound Poisson process to model disease-related events as the primary unit of analysis for SCD rather than analyzing data of individuals in a case/non-case fashion. Such a model enables us to detect geographical clusters of events when individuals in a population may present multiple events (e.g., emergency department (ED) visits, post-emergency physician/practitioner visits) related to a disease or disorder diagnosis. Events generated by the same individual should be deemed correlated by a probability distribution. In Section 2, we introduce the notation used throughout the paper and our compound Poisson model. The test statistic is introduced and the procedure to assess its p-value is outlined. In Section 3, we describe our administrative data and present a case study to illustrate our methodology. Simulated data sets are analyzed for further illustrative purposes in Section 4. Some concluding remarks of the topic and future research ideas are organized in Section 5.

2. Methodology

We assume administrative health data can be segregated into I non-overlapping geographic sub-divisions (sometimes called subregions or cells). Suppose each subregion is characterized by a centroid. A zone Z, which is defined by a circular spatial scan window of radius r and its centre at the coordinate of a centroid, consists of only and all individuals in those subregions whose centroids lie inside the circle [2].

For this type of two dimensional scan test, we may choose an upper bound r_i^*, $i = 1, \cdots, I$, on the radius of the circular scan window such that the population size of any zone defined by the window centered at centroid i does not exceed β percent of the total population in the study region. As indicated by Kulldorff and Nagarwalla [2], the choice of the upper bound on r_i should be decided prior to analysis. All test zones can be generated by combining nearest subregions with subregion i by varying the radius of the defining circle from 0 to r_i^* for each i. Clearly when $r_i = 0$ for all i, each zone coincides with a single subregion. Zones with such a definition have irregular geographic boundaries that depend on the size and shape of those subregions whose centroids lie inside the spatial scan window.

Consider the hypothetical seven-cell region in Figure 1 as an illustration. Circular scan windows of various radii are centered at the upper north centroid and the centroid at the southwest corner. Starting from each centroid, a new test zone

is formed when a new neighboring centroid is enclosed by the scan window. The value of r_i^* varies for each centroid i depending on the population size of cell i and of its nearby neighbors and the chosen β.

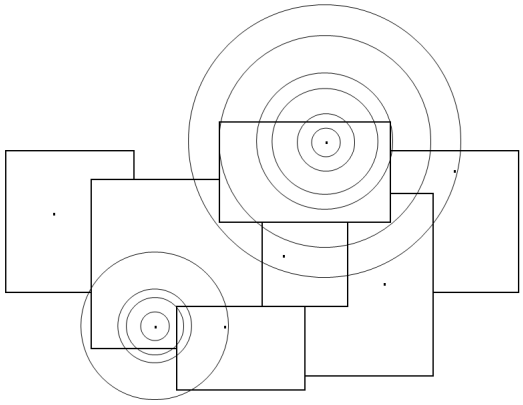

Figure 1. hypothetical seven-cell region.

2.1. Notation, Compound Poisson Model and the Spatial Scan Test

Let Z be a collection of distinct subregions (administrative areas) in geographic region R. For each predetermined Z $\subset$ R, let Z$'$ $\subset$ R denote the compliment of Z, such that Z $\cap$ Z$'$ $= \emptyset$ and Z $\cup$ Z$'$ $=$ R. Let there be a total of I non-overlapping and contiguous subregions S_i (such as districts of land space), $i = 1, \cdots, I$, in R (such as a state or a province).

Let the random variable C_{ik}, with observed value c_{ik}, be the number of individuals observed with k events in subregion S_i ($k \in \mathbb{N}^+$, $i = 1, \ldots, I$). Let the random variable $C_i = \sum_{k \in \mathbb{N}^+} C_{ik}$, with observed value c_i, be the total number of individuals with at least one event in subregion S_i. With the definition of a test zone in the previous section, the random variable $C_{\mathrm{Z}} = \sum_{\substack{i=1 \\ \mathrm{S}_i \in \mathrm{Z}}}^{I} C_i$, with observed value c_{Z}, is the number of individuals (cases) with at least one event in Z, and similarly for outside the test zone, $C_{\mathrm{Z}'} = \sum_{\substack{i=1 \\ \mathrm{S}_i \notin \mathrm{Z}}}^{I} C_i$, with observed value $c_{\mathrm{Z}'}$. Let $C = C_{\mathrm{Z}} + C_{\mathrm{Z}'}$ denote the total number of cases in R. We wish to detect zones in R that have higher than expected numbers of events. The spatial scan statistic we propose is based on a likelihood ratio test which is in the same spirit as in [2].

In this paper, we assume the population size of subregion S_i can be measured and is denoted by n_i for $i = 1, \ldots, I$, and that $C_i \sim \mathrm{POI}(\lambda_i n_i)$ where $\lambda_i > 0$ are standardized Poisson intensities. We consider that within a time period, say fiscal year, individuals will only have event(s) within the subregion of their residence. Of the individuals with at least one event, let the random variable $X_{i\ell}$ denote the number of event(s) incurred by the ℓth individual in subregion S_i, $\ell = 1, \ldots, C_i$. The density distribution of the random variable $X_{i\ell}$ can be arbitrary depending on the context, however, we assume that $X_{i\ell}$ is discrete and follows a zero-truncated Poisson distribution with density

$$\Pr(X_{i\ell} = x; \theta_i) = \begin{cases} Q(x;\theta_i) = \dfrac{\theta_i^x}{x!(e^{\theta_i} - 1)} & \text{for } x = 1, 2, \cdots \\ 0 & \text{elsewhere} \end{cases} \tag{1}$$

with mean $\theta_i e^{\theta_i}/(e^{\theta_i} - 1)$ where $\theta_i > 0$.

The total number of events from the population of subregion S_i can be written as

$$U_i = \sum_{\ell=1}^{C_i} X_{i\ell}$$

where it is reasonable to assume that C_i is independent of $X_{i\ell}$ for $\ell = 1, \ldots, C_i$ and $i = 1, \ldots, I$. For example, C_i can be used to represent, over a fixed period of time, the number of people having respiratory symptoms in subregion i and $X_{i\ell}$

the number of times the ℓth individual of subregion i visited for hospital emergency service with such symptoms. With this formulation, U_i is a compound Poisson random variable. Panjer [18] discovered that the probabilities $\Pr(U_i = u_i)$ can be obtained using a simple recursion formula

$$\begin{aligned} \Pr(U_i = 0) &= e^{-\lambda_i n_i}, \\ \Pr(U_i = u_i) &= \frac{\lambda_i n_i}{u_i} \sum_{x=1}^{u_i} xQ(x;\theta_i) \Pr(U_i = u_i - x) \quad u_i = 1, 2, \cdots, \quad i = 1, \cdots, I. \end{aligned} \tag{2}$$

With $Q(x;\theta_i)$ known, the likelihood function for the compound Poisson model can be written as

$$L(\lambda_i, \theta_i) \propto \prod_{\substack{i=1 \\ C_i>0}}^{I} \frac{\lambda_i n_i}{u_i} \sum_{x=1}^{u_i} xQ(x,\theta_i) \Pr(U_i = u_i - x) \tag{3}$$

since our main interest is on test zones having at least one event.

If we assume under the null hypothesis $\mathrm{H_0}$: $\lambda_i = \lambda$ and $\theta_i = \theta$ for all i, the likelihood becomes

$$L(\hat{\lambda}, \hat{\theta}) = \prod_{\substack{i=1 \\ C_i=0}}^{I} e^{-\hat{\lambda} n_i} \prod_{\substack{i=1 \\ C_i>0}}^{I} \frac{\hat{\lambda} n_i}{u_i} \sum_{x=1}^{u_i} xQ(x;\hat{\theta}) \Pr(U_i = u_i - x) \tag{4}$$

where $\hat{\lambda}$ and $\hat{\theta}$ are maximum likelihood (ML) estimates of λ and θ, respectively. If we assume that under the alternative hypothesis $\mathrm{H_a}$: $\lambda_i = \lambda_{\mathrm{Z}}$ and $\theta_i = \mu$ for $\mathrm{S}_i \in \mathrm{Z}$, $\lambda_i = \lambda_{\mathrm{Z}'}$ and $\theta_i = \nu$ for $\mathrm{S}_i \notin \mathrm{Z}$, and $\lambda_{\mathrm{Z}} \mu e^{\mu}/(e^{\mu} - 1) > \lambda_{\mathrm{Z}'} \nu e^{\nu}/(e^{\nu} - 1)$, the probabilities $\Pr(U_i = u_i)$ can be obtained using separate recursion equations similar to (2) for $\mathrm{S}_i \in \mathrm{Z}$ and $\mathrm{S}_i \notin \mathrm{Z}$, $i = 1, \ldots, I$. Note that we can interpret the inequality in the alternative hypothesis statement as: per fixed population size, the expected number of events incurred inside a test zone is higher than that outside the zone. Conditional on $U_i = u_i$ for $i = 1, \ldots, I$, the likelihood under the alternative becomes $L(\mathrm{Z}, \hat{\lambda}_{\mathrm{Z}}, \hat{\lambda}_{\mathrm{Z}'}, \hat{\mu}, \hat{\nu})$ having the following expression

$$\prod_{\substack{i=1 \\ \mathrm{S}_i \in \mathrm{Z}, C_i=0}}^{I} e^{-\hat{\lambda}_{\mathrm{Z}} n_i} \prod_{\substack{i=1 \\ \mathrm{S}_i \in \mathrm{Z}, C_i>0}}^{I} \frac{\hat{\lambda}_{\mathrm{Z}} n_i}{u_i} \sum_{x=1}^{u_i} xQ(x;\hat{\mu}) \Pr(U_i = u_i - x) \prod_{\substack{i=1 \\ \mathrm{S}_i \notin \mathrm{Z}, C_i=0}}^{I} e^{-\hat{\lambda}_{\mathrm{Z}'} n_i} \prod_{\substack{i=1 \\ \mathrm{S}_i \notin \mathrm{Z}, C_i>0}}^{I} \frac{\hat{\lambda}_{\mathrm{Z}'} n_i}{u_i} \sum_{x=1}^{u_i} xQ(x;\hat{\nu}) \Pr(U_i = u_i - x) \tag{5}$$

where $\hat{\lambda}_{\mathrm{Z}}, \hat{\lambda}_{\mathrm{Z}'}, \hat{\mu}$ and $\hat{\nu}$ are the ML estimates of their respective denoted parameters.

2.2. Likelihood Ratio Test Statistic

We choose the likelihood ratio test statistic to be

$$\eta = \frac{\max\limits_{\mathrm{Z} \subset \mathrm{R}} L_{\hat{\mu},\hat{\nu}}(\mathrm{Z}, \hat{\lambda}_{\mathrm{Z}}, \hat{\lambda}_{\mathrm{Z}'})}{L_{\hat{\theta}}(\hat{\lambda})} \quad \text{or} \quad \eta = \max_{\mathrm{Z} \subset \mathrm{R}} \log \frac{L_{\hat{\mu},\hat{\nu}}(\mathrm{Z}, \hat{\lambda}_{\mathrm{Z}}, \hat{\lambda}_{\mathrm{Z}'})}{L_{\hat{\theta}}(\hat{\lambda})} \tag{6}$$

with $\hat{\lambda}$ denoting the ML estimate of λ under the null hypothesis. We suggest that $\hat{\theta}$ of $Q(x;\theta)$ in (4) be obtained first also by the maximum likelihood method. Since the only variables in (3) are n_i and u_i, various combinations of $x_{i\ell}$ for $\ell = 1, \cdots, c_i$ may contribute to the same set of u_i, $i = 1, \cdots, I$, to yield the same $\hat{\lambda}$ and $\hat{\theta}$ for (4) if they are jointly estimated by the ML method. In direct joint ML estimation of λ and θ using equation (3), information on the individual event numbers is actually not used in estimating θ. For instance, if we observe that subregion i has five cases ($C_i = 5$), then both $\{X_{i1} = 1, X_{i2} = 2, X_{i3} = 3, X_{i4} = 4, X_{i5} = 5\}$ and $\{X_{i1} = 3, X_{i2} = 3, X_{i3} = 3, X_{i4} = 3, X_{i5} = 3\}$ give $U_i = 15$ and will yield the same $\hat{\lambda}$ and $\hat{\theta}$ if they were jointly estimated from U_i and (4). This parameter identifiability issue is eliminated if $\hat{\theta}$ is estimated from $X_{i\ell}$ and $Q(x,\theta)$. Under the null hypothesis, for example, we choose $\hat{\theta}$ to be the ML

estimate which maximizes

$$\log \prod_{i=1}^{I} \prod_{\ell=1}^{C_i} Q(X_{i\ell} = x; \hat{\theta}).$$

If each and every case in every region only has one event, the spatial scan test using the Poisson model will be more appropriate.

Estimating $\hat{\theta}$ first, based on $X_{i\ell}$, then substituting $\hat{\theta}$ into (3) allows us to obtain a ML estimate of λ based on $L_{\hat{\theta}}(\lambda)$. $L_{\hat{\theta}}(\hat{\lambda})$ can then be evaluated for the likelihood ratio test statistic. In a similar fashion but under the alternative hypothesis, for each test zone Z and Z′, $\hat{\lambda}_{\mathrm{Z}}$ and $\hat{\lambda}_{\mathrm{Z}'}$ are joint ML estimates based on (5) after obtaining $\hat{\mu}$ and $\hat{\nu}$ first for $Q(x, \mu)$ and $Q(x, \nu)$ from $\mathrm{S}_i \in \mathrm{Z}$ and $\mathrm{S}_i \notin \mathrm{Z}$, respectively. Note that in practice, we only consider η from zones such that the condition

$$\frac{\hat{\lambda}_{\mathrm{Z}}\hat{\mu}e^{\hat{\mu}}}{(e^{\hat{\mu}}-1)} > \frac{\hat{\lambda}_{\mathrm{Z}'}\hat{\nu}e^{\hat{\nu}}}{(e^{\hat{\nu}}-1)} \quad \text{or} \quad \phi = \frac{\hat{\lambda}_{\mathrm{Z}}\hat{\mu}e^{\hat{\mu}}(e^{\hat{\nu}}-1)}{\hat{\lambda}_{\mathrm{Z}'}\hat{\nu}e^{\hat{\nu}}(e^{\hat{\mu}}-1)} > 1 \tag{7}$$

is also satisfied since our interest is in finding regions of high expected number of events.

As for the spatial scan statistic introduced by Kulldorff [19], the exact distribution of η in an analytical form is very difficult to obtain. Monte Carlo simulation is to be employed to assess the significance of an observed value of η under the null hypothesis by taking the following steps. For all simulation procedures, we condition on $C = c$ and the number of events of the jth case in the study area regardless of subregion, $X_j = x_j$, for $j = 1, \ldots, c$. Note that given an individual generates at least one event in R, the probability that such an individual belongs to subregion S_i is n_i/n under the null assumptions.

1. Conditioning on $C = c$ and $X_j = x_j$, sample randomly a subregion ID for each x_j, $j = 1, \cdots, c$. The sampling distribution has weights n_i/n for $i = 1, \cdots, I$. Depending on the generated subregion ID for each individual, variables C_i and $X_{i\ell} = x_{i\ell}$, hence U_i, are generated for $\ell = 1, \ldots, C_i$ and $i = 1, \ldots, I$.
2. Calculate the test statistic η as defined in (6).
3. Repeat steps 1 and 2 for 999 trials and record test statistic of each simulation trial.
4. Rank the 999 simulated likelihood ratio statistics and the observed statistic η from the data.

Note that in our model, the likelihood $L_{\hat{\theta}}(\hat{\lambda})$ is no longer a constant under the null hypothesis over each simulation trial as it would be in [2] and [19], because (3) depends on $U_i = u_i$ for $i = 1, \cdots, I$ which are not fixed in the simulation trials. Therefore, the numerator and the denominator of (6) need to be computed in step 2 of each simulation trial. The hypothesis test can be considered significant at 1000α per cent level if the value of the observed η calculated from data is among the 1000α (an integer) highest of these 1000 ranked statistics in step 4. A significant test indicates that the collection of subregions which yields the observed η in the spatial scan test is the most likely cluster having higher expected numbers of events per fixed population size. Other zones which have nonoverlapping subregions with the most likely cluster also have high values of the test statistic under condition (7) should be examined for possibility of being secondary clusters.

3. Application to Emergency Data

We illustrate our spatial scan on emergency department (ED) presentations by children and youth (age <18 years) for substance abuse in the western Canadian province of Alberta during six fiscal years (April 1, 2002, to March 31, 2008). The data are extracted from population-based, provincial administrative data sets held at Alberta Health as part of a larger study investigating ED presentations for mental health conditions [21]. Each ED presentation during the study period is considered to be an event made by an individual and is tied back to the sRHA code of patient residence. A case is defined as an individual who had at least one ED presentation for substance abuse in Alberta during the study period. We examine each fiscal year separately using both our new spatial scan for events and two analyses using Kulldorff's spatial scan

Table 1. *Subregional Population Size by Fiscal Years*

	sRHA													
Fiscal Year	1	2	3	4	5	6	7	···	65	66	67	68	69	70
2002/2003	3743	7466	17573	8574	5357	19191	6692	···	6929	20083	2809	3357	3643	13280
2003/2004	3591	7372	17717	8473	5338	19288	6687	···	6951	20357	2772	3477	3701	13529
2004/2005	3471	7223	17783	8394	5417	19397	6686	···	6743	20740	2724	3535	3780	13756
2005/2006	3365	7129	17745	8376	5341	19432	6819	···	6690	21440	2712	3722	3740	13671
2006/2007	3379	7253	18362	8388	5567	19906	6965	···	6588	22444	2673	3895	3770	14250
2007/2008	3272	7392	18970	8535	5690	20336	6937	···	6596	23131	2592	4024	3839	14807

software SaTScan [22], where analyses are performed on cases only and on events as if events were individual cases (i.e., ignoring correlation).

There are $I = 70$ distinct subregional health authorities (sRHA) in the province of Alberta, Canada and we use these as the subregions for analysis. The population size of each sRHA is partially reported in Table 1 by fiscal year. In Table 2, the total number of cases and ED presentations related to substance abuse in each sRHA are partially reported. Note that for some years and sRHAs, the cases and events are zero meaning that these sRHAs would not be included individually as test zones.

Table 2. *Cases and Events by Fiscal Year and SubRegional Health Authority*

	sRHA											
Fiscal Year	1	2	3	4	5	6	7	8	···	68	69	70
2002/2003	7 (8)	12 (13)	20 (20)	7 (8)	– (–)	24 (27)	9 (11)	7 (7)	···	0 (0)	– (–)	18 (21)
2003/2004	– (–)	20 (20)	22 (22)	– (–)	– (–)	24 (26)	12 (13)	15 (16)	···	0 (0)	7 (7)	18 (18)
2004/2005	6 (7)	16 (21)	27 (28)	6 (6)	– (–)	29 (30)	20 (23)	9 (9)	···	0 (0)	7 (9)	29 (30)
2005/2006	10 (10)	18 (18)	20 (21)	6 (6)	– (–)	36 (39)	9 (10)	9 (9)	···	0 (0)	0 (0)	33 (37)
2006/2007	12 (13)	24 (25)	23 (25)	– (–)	15 (15)	35 (35)	18 (19)	19 (19)	···	0 (0)	– (–)	18 (21)
2007/2008	11 (13)	22 (24)	30 (34)	– (–)	– (–)	34 (37)	10 (10)	19 (20)	···	0 (0)	11 (12)	27 (32)

– denotes cell counts are <6 and are suppressed to ensure confidentiality

The test zones are created with up to 7% of the study population to allow for between 250 and 300 zones for each fiscal year. Alberta has a sparse population (around 800,000 children and youth) for its geographic area, and combining multiple sRHAs into a test zone could lead to large geographic regions which would likely not be clusters. The most likely clusters for each fiscal year are provided in Table 3 and Figure 2, and the detailed results for fiscal year 2005/2006 are provided in Table 4.

Table 3. *Retrospective Analysis, Compound Poisson Model,* $\beta = 7$

		statistical summary			
Fiscal Year	sRHA(s)	case (event) count	log L.R. Stat.	ϕ	p-value
2002/2003	{60,61,62,70}	97 (102)	13.44	1.54	<0.001
2003/2004	{47,50}	75 (76)	12.69	1.02	<0.001
2004/2005	{61,62,70}	114 (125)	15.13	1.80	<0.001
2005/2006	{61,62,70}	119 (131)	15.90	1.54	<0.001
2006/2007	{35,36,37,38,39,40,52,62}	85 (91)	19.33	1.04	<0.001
2007/2008	{61,62,70}	114 (128)	15.58	1.81	<0.001

Table 4 shows the sRHAs in each test zone and the associated log likelihood ratios (L.L.R.) and ϕ values. The zone with the highest observed value of the log likelihood test statistic (obs = 15.90) and the observed value $\phi = 1.54 > 1$ consists of sRHAs 61, 62 and 70. The compound Poisson model suggests that these subregions, with 38, 56 and 37 events, respectively, and population sizes of 12031, 17365 and 13671, respectively, together form the most likely zone of an event cluster. The three subregions have over one and half times the expected number of ED visits than what is expected in the other subregions, per capita. Through a simulation of 999 trials, the observed statistic $\eta = 15.90$ ranked 1st of the

Table 4. *Mental/Behavioural disorder due to psychoactive substance abuse, 2005/2006, β=7*

sRHA(s)	L.L.R.	ϕ	sRHA(s)	L.L.R.	ϕ	sRHA(s)	L.L.R.	ϕ
{1}	0.3792	1.4961	{10,13}	-18.7081	4.1703	{63,64}	-1.2480	1.1915
{1,2}	-0.8030	1.4058	{11}	-5.1287	3.4624	{63,64,67}	2.1757	1.3345
{1,2,26}	1.2355	1.7844	{8,11,15}	-14.1452	2.4367	{63,64,67,68}	-1.2119	0.9676
{1,2,3,26}	2.1435	1.2537	{12}	-6.2859	3.1230	{63,64,65,67,68}	-3.1185	1.1823
{1,2,3,4,26}	0.1292	1.0913	{9,12,15}	-7.4701	1.7090	{63,64,65,66,67,68}	-1.4166	1.1950
{1,2,3,4,25,26}	-0.3605	1.1090	{13}	-7.5173	3.8379	{64}	-0.0470	1.8782
{2}	0.0257	1.2444	{10,13,17}	-35.1713	6.6892	{63,64,65}	-1.2548	1.3767
{2,3}	0.7784	0.7484	{14}	-4.8699	2.2511	{63,64,65,68}	-6.4589	1.1148
{2,3,4}	-1.2600	0.7475	{14,15}	-9.2553	2.5372	{61,63,64,65,68}	-1.7907	1.2677
{1,2,3,4}	-1.1585	0.8824	{11,14,15}	-14.3455	2.7709	{61,63,64,65,66,68}	3.7669	1.3190
{1,2,3,4,5}	1.0797	0.8882	{11,14,15,18}	-23.4622	3.5059	{65}	0.8827	1.4790
{1,2,3,4,5,26}	-0.2099	1.0999	{15}	-4.3171	3.1336	{65,66}	4.2326	1.4450
{3}	1.0474	0.5509	{11,12,14,15}	-20.5659	2.8077	{64,65,66}	4.7362	1.5995
{3,4}	-16.3610	6.2500	{16}	0.3704	2.0443	{60,64,65,66}	5.5858	1.3979
{2,3,4,5}	0.4944	0.8173	{16,17}	0.6109	0.8895	{59,60,64,65,66}	8.1739	1.3964
{1,2,3,4,5,25}	3.0463	0.8831	{16,17,19}	1.6716	0.8023	{66}	6.5460	0.9536
{4}	-0.5443	2.0968	{16,17,19,20}	0.7427	0.6089	{59,65,66}	8.2643	1.4404
{3,4,5}	-14.7075	4.9071	{16,17,18,19,20}	-0.1387	0.6015	{59,63,65,66}	5.2423	1.3564
{1,2,3,4,5,7,25}	3.3243	0.8552		·		{59,63,64,65,66}	4.9958	1.4052
{5}	1.3630	2.2911		·		{67}	11.1926	0.1808
{4,5}	0.8896	2.1473		·		{67,68}	5.8120	0.0752
{3,4,5,7}	3.6686	0.6501	{59}	3.7755	1.2434	{63,67,68}	-1.0549	1.6814
{3,4,5,6,7}	0.5933	0.8889	{28,59}	3.1805	1.5258	{63,64,67,68,69}	-6.0829	0.7282
{6}	5.0360	0.9162	{27,28,59}	1.3019	1.7017	{63,64,65,67,68,69}	-7.7689	0.9810
{6,7}	-0.2419	1.2656	{27,28,56,59}	2.6235	1.1883	{63,64,65,66,67,68,69}	-4.7750	1.0609
{5,6,7}	0.3822	1.1106	{27,28,56,59,65}	3.7688	1.2815	{64,68}	-5.9884	1.0118
{4,5,6,7}	2.6187	1.0120	{60}	0.0185	0.8505	{64,67,68}	0.6349	0.9744
{4,5,6,7,25}	4.9580	0.9771	{56,60}	2.0780	0.7918	{61,63,64,67,68,69}	-5.3024	1.0892
{7}	-0.2183	0.6879	{56,57,60}	1.1048	0.9335	{61,63,64,65,67,68,69}	-6.0414	1.2084
{6,7,25}	0.1453	1.1364	{41,56,57,60}	2.9072	0.8734	{61,63,64,65,67,68,69,70}	-4.5948	1.2939
{5,6,7,25}	2.3659	1.1233	{61}	4.8650	1.4525	{69,70}	-0.9448	0.9806
{5,6,7,25,34}	3.7096	1.0565	{61,70}	9.0281	1.3370	{68,69,70}	-6.4182	0.8205
{4,5,6,7,25,34}	6.1627	0.9392	{***61,62,70***}	***15.9000***	***1.5399***	{61,68,69,70}	-1.2853	1.0493
{8}	-4.5483	1.7352	{60,61,62,70}	13.2156	1.6563	{61,64,68,69,70}	-1.4820	1.1560
{8,11}	-9.7680	2.2830	{62}	9.1143	1.4841	{61,64,67,68,69,70}	-3.1589	1.2901
{8,9,11}	-6.4775	1.4241	{36,62}	-0.3287	2.0042	{61,62,64,67,68,69,70}	1.7970	1.5509
{9}	3.5985	0.4037	{36,37,62}	-1.6605	1.6596	{70}	4.1170	1.2010
{9,12}	-2.9741	1.3460	{36,37,52,58,62}	2.1736	1.4539	{61,69,70}	2.5805	1.5026
{8,9,12}	-7.6109	1.4367	{36,37,52,57,58,62}	1.5499	1.3415	{61,62,69,70}	12.1263	1.3121
{10}	-11.1946	4.6576	{63}	-0.7186	0.7597	{37,61,62,69,70}	0.7680	1.5187

the top 50 largest, identifying the cluster as statistically significant at the 5% significance level. On the other hand, zones associated with high log likelihood ratios but low $\phi < 1$ values are areas with fewer expected number of event counts than what is expected in the other regions and is not of interest for our particular purposes. The sRHAs identified in each year differ, although sRHAs 61, 62 and 70 together appear in the most likely cluster for four out of six consecutive years and would be the most likely cluster in a spatial-temporal analysis with a temporal window of one year. These sRHAs together had at least one and a half times the expected number of ED visits than what is expected in the other subregions.

Table 5 shows the most likely clusters identified with the traditional spatial scan [19] for each year separately based on the discrete Poisson model. The first analysis only examines cases whilst the second analyses performs the test on the events (i.e., treating them as if they are cases and ignoring correlation). Similar zones are identified for fiscal years 2004/2005 and 2005/2006, however, the analysis based on events has an additional sRHA 69 in fiscal year 2004/2005. For other fiscal years, the analysis based on cases and events yield quite different findings from the analyses based on our compound Poisson model.

All the scripts used for data analyses based on our compound Poisson model are implemented in Matlab [20]. The amount of time it takes to run a complete spatial scan test for 2002/2003 and 2003/2004 fiscal year data by our compound

Table 5. *Retrospective Analysis, Discrete Poisson Model, β=7*

	analysis based on case				analysis based on event			
Fiscal Year	sRHA(s)	count	log L.R. Stat.	p-value	sRHA(s)	count	log L.R. Stat.	p-value
2002/2003	{46}	48	12.00	4.2×10^{-5}	{46}	55	15.77	5.3×10^{-7}
2003/2004	{28,29,30,53,56}	73	14.69	5.0×10^{-6}	{28,29,30,53,56}	78	19.72	1.5×10^{-8}
2004/2005	{61,62,70}	114	18.32	4.0×10^{-8}	{61,62,69,70}	134	21.67	2.6×10^{-9}
2005/2006	{61,62,70}	119	18.88	6.9×10^{-8}	{61,62,70}	131	18.80	7.5×10^{-8}
2006/2007	{61}	58	27.31	8.3×10^{-12}	{61}	72	41.47	1.0×10^{-17}
2007/2008	{44}	49	22.18	7.9×10^{-10}	{44}	59	30.44	2.7×10^{-13}

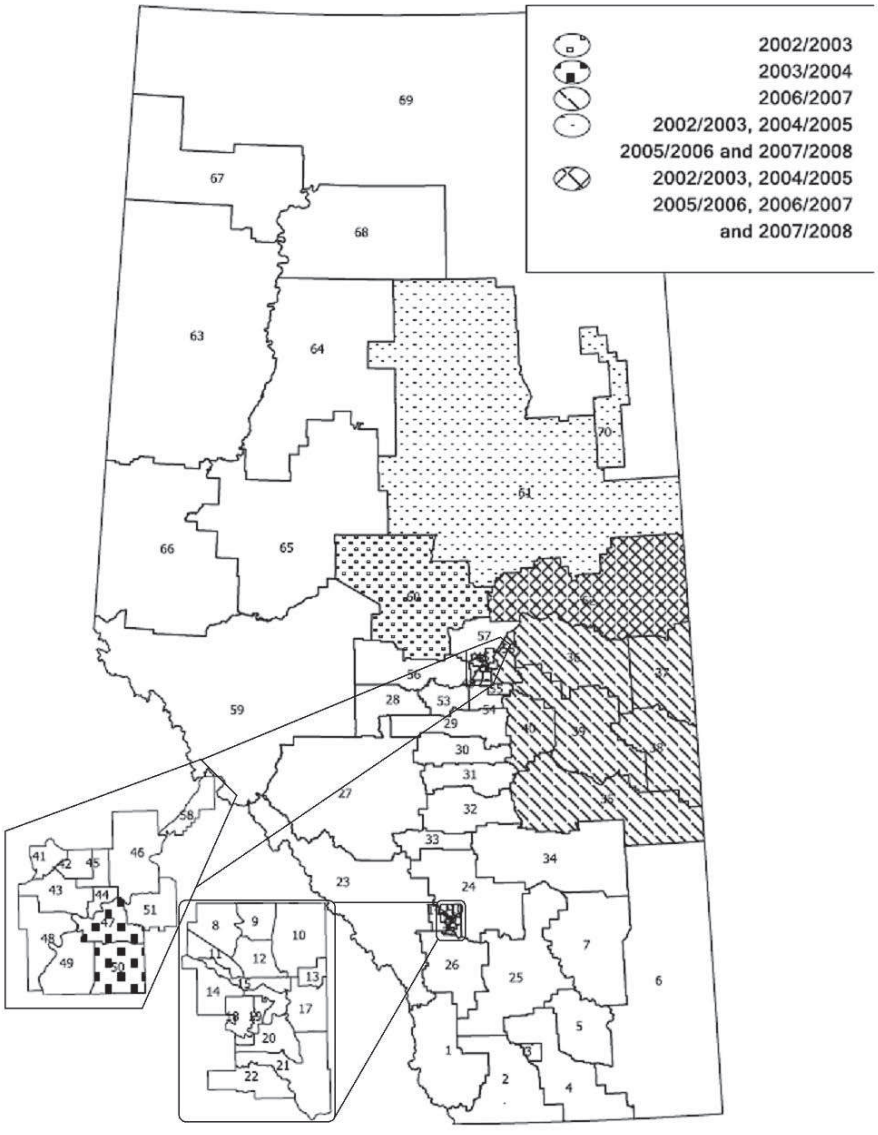

Figure 2. Possible clusters of events in Alberta, Canada.

Poisson model is 2.9 and 4.5 hours, respectively, on a desktop computer equipped with single quad-core Intel i7 processor running at 3.5 GHz. It takes between 3 to 5 hours to complete tests on the data of other fiscal years. The computing time is dictated by the programming language used. The computing time may be improved dramatically if implemented in C/C++ by a statistical programmer, for example. The computing time to complete the spatial scan test by the discrete Poisson model using SaTScan is within seconds for the aggregated data of each fiscal year. SaTScan also allows the choice of

using circular or elliptical window shapes, as well as the option of using the Isotonic Spatial Scan Statistic which is not yet available in our approach.

4. Simulation Study

To further our understanding of how the new spatial scan may differ from analyses using Kulldorff's [19] spatial scan, we generate random datasets based on Alberta's geography. For simplicity, we re-label the 70 sRHAs as subregions $S_1, \cdots, S_{70}$ in the province and the regional population sizes are taken from the 2005/2006 fiscal year. We first generate the number of cases $C_i \sim \text{POI}(\lambda_i n_i)$ of each subregion where $\lambda_i = 0.0008$ for all i, except we choose $\lambda_i = 0.003$ for subregions S_{42}, S_{44}, S_{45}, S_{61} and S_{62}, and set $\lambda_{70} = 0.0025$ for subregion S_{70}. The next step is to generate the number of events for each case by $Q(x;\theta)$ in (1). The distribution $Q(x;\theta = 1.5)$ is used to generate the number of events for each case in each and every subregion, except $Q(x;\theta = 3)$ is used for S_{42}, S_{44}, S_{45}, S_{61} and S_{70}. To give a better idea of the two densities, we tabulate the first 10 probability values of both distributions in Table 6

Table 6. *Zero-truncated Poisson Density*

	x									
	1	2	3	4	5	6	7	8	9	10
$Q(x;\theta = 1.5)$	0.4308	0.3231	0.1616	0.0606	0.0182	0.0045	0.0010	0.0002	0.0000	0.0000
$Q(x;\theta = 3.0)$	0.1572	0.2358	0.2358	0.1768	0.1061	0.0531	0.0227	0.0085	0.0028	0.0009

In our past experience with emergency visit data sets, patients tended to make fewer than 10 visits per year, but some could make 40 or more visits. The distribution of events may not be simple and motivates the examination of the simple Poisson model in SaTScan [22] and our compound Poisson model where cases and events are generated by a mixture.

We choose three subregions S_{14}, S_{15} and S_{70} to generate an additional number of cases, $C_i^* \sim \text{POI}(\lambda_i^* n_i)$ where $\lambda_{14}^* = 0.0003$, $\lambda_{15}^* = 0.0004$, and $\lambda_{70}^* = 0.0005$. Overall, the rate of the number of cases generated in S_{70} is $\lambda_{70} + \lambda_{70}^* = 0.003$, which is the same as those of S_i for $i = 42, 44, 45, 61$ and 62. This rate is at least three times the rate of the other subregions. For each of the additional C_i^* cases, the following custom discrete distribution

$$g(x;\vartheta) = \begin{cases} \left(1 - \sum_{x=26}^{50} \frac{x^2\vartheta}{100}\right)/25 & \text{for } x = 1, \cdots, 25 \\ \frac{x^2\vartheta}{100} & \text{for } x = 26, \cdots, 50 \\ 0 & \text{elsewhere.} \end{cases}$$

with $\vartheta = 1/400$ is used to generate the number of events. This distribution allows much higher probability of generating 26 to 50 events per individual. In this setup, higher event rates should easily be identified for the three zones $\{S_{14}, S_{15}\}$, $\{S_{41}, S_{42}, S_{45}\}$ and $\{S_{61}, S_{62}, S_{70}\}$ due to geographical distance, and the three major zones are far apart from each other without overlapping subregions. In particular, since S_{70} has a slightly higher mean number of events per case, our conjecture is that the zone $\{S_{61}, S_{62}, S_{70}\}$ shall be more likely to be detected as the most likely cluster of events.

We ran a simulation study of 100 experiments by generating data sets as described above. A sample data set is given with the population (Pop.) size of each sRHA in 2005/2006 in Table 7. We first ran a purely spatial analysis and choose a Poisson discrete probability model in SaTScan [22] for case data only. The maximum spatial cluster size is set to be 7% and the spatial window shape is set to be circular. We then analyze the event data separately by the same procedure using SaTScan [22] and by our compound Poisson model. The identified most likely cluster and the number of times it has been identified are tabulated in Table 8.

Table 7. *A Sample of Generated Data Set*

Dataset 15											
S_i	Pop.	c_i	u_i	S_i	Pop.	c_i	u_i	S_i	Pop.	c_i	u_i
S_1	3365	4	11	S_{25}	5045	2	3	S_{49}	23065	23	50
S_2	7129	4	13	S_{26}	13868	7	15	S_{50}	28609	26	61
S_3	17745	12	20	S_{27}	5627	3	6	S_{51}	15305	9	14
S_4	8376	6	11	S_{28}	4358	3	4	S_{52}	6304	6	10
S_5	5341	4	11	S_{29}	11065	9	20	S_{53}	2224	1	3
S_6	19432	16	24	S_{30}	4427	4	8	S_{54}	5114	5	7
S_7	6819	5	7	S_{31}	8524	2	6	S_{55}	3695	2	3
S_8	22409	13	21	S_{32}	30344	33	70	S_{56}	20010	16	34
S_9	19061	14	26	S_{33}	3998	3	6	S_{57}	9252	6	13
S_{10}	23234	19	38	S_{34}	6556	2	2	S_{58}	4064	6	12
S_{11}	11387	10	19	S_{35}	4812	6	9	S_{59}	11592	9	15
S_{12}	15513	10	17	S_{36}	5311	2	5	S_{60}	10417	5	7
S_{13}	15286	10	26	S_{37}	8053	2	4	S_{61}	11998	29	74
S_{14}	17383	16	74	S_{38}	3943	4	6	S_{62}	17471	50	148
S_{15}	9954	14	118	S_{39}	5296	4	6	S_{63}	7342	6	10
S_{16}	5981	5	9	S_{40}	5937	6	11	S_{64}	4474	5	10
S_{17}	15704	10	18	S_{41}	14575	16	29	S_{65}	6690	4	6
S_{18}	10067	6	20	S_{42}	12031	42	152	S_{66}	21440	13	20
S_{19}	6946	7	9	S_{43}	14134	12	20	S_{67}	2712	1	2
S_{20}	19065	20	31	S_{44}	13671	49	163	S_{68}	3722	7	14
S_{21}	22838	10	21	S_{45}	17365	40	128	S_{69}	3740	2	3
S_{22}	20234	24	44	S_{46}	19415	13	27	S_{70}	14250	44	278
S_{23}	11152	7	16	S_{47}	14564	4	11				
S_{24}	26786	29	42	S_{48}	24188	15	26				

Table 8. *Number of Detection in Simulation Study, β=7*

	Poisson Model		Compound Poisson Model
Cluster	Analyzing Cases	Analyzing Events	Analyzing Cases and Events
$\{S_{61},S_{62},S_{70}\}$	55	59	83
$\{S_{42},S_{44},S_{45}\}$	42	4	4
$\{S_{14},S_{15}\}$	—	15	8
$\{S_{61},S_{62},S_{69},S_{70}\}$	3	—	—
$\{S_{14}\}$	—	1	—
$\{S_{61},S_{70}\}$	—	5	4
$\{S_{70}\}$	—	16	1

— denotes values which are not available

The two zones $\{S_{61}, S_{62}, S_{70}\}$ and $\{S_{41}, S_{42}, S_{45}\}$ have the same overall rate $\lambda = 0.003$, and when analyzing only cases with a Poisson model, they have similar chance of being identified as the most likely case cluster. In this study, $\{S_{61}, S_{62}, S_{70}\}$ are detected more frequently (55 times versus 42 for $\{S_{41}, S_{42}, S_{45}\}$), the margin is small.

When analyzing only events using the Poisson model which does not take into account the intra-person correlation of data, the spatial scan test of SaTScan [22] identified $\{S_{61}, S_{62}, S_{70}\}$ as the most likely cluster 59 times which is considerably more frequent than other identified event clusters. This confirms with our initial conjecture that in our setup, $\{S_{61}, S_{62}, S_{70}\}$ should be more probable of being a cluster. If we take into account the intra-person correlations of events generated by each case, our compound Poisson model is able to detect $\{S_{61}, S_{62}, S_{70}\}$ as the most likely cluster 83 times which has a considerable higher success rate under the setup of this simulation study.

5. Discussion

Spatial cluster detection tests usually attempt to identify geographic regions with higher than expected numbers of incident or prevalent cases of disease or illness. We are interested in detecting geographic regions with higher than expected

numbers of events related to disease or illness, with the particular feature that individual cases may have multiple disease-related events (i.e., correlated data). We have treated the number of events as a compound Poisson random variable and proposed a spatial scan statistic for compound Poisson data. To permit parameter identifiability, an estimate of the zero-truncated Poisson parameter is obtained and a likelihood ratio test statistic is developed based on the parameter estimates inside and outside the tested zone. An inequality based on mean number of events was used to help identify zones with higher, rather than lower, numbers of events expected. Monte Carlo simulations are conducted to assess the significance of zones.

We applied our method to substance abuse presentations by children and youth to Alberta emergency departments during a six-year period and compared the new approach with two applications of the traditional spatial scan: applying the spatial scan to case counts and applying spatial scan to event counts as if the event counts were independent data. We adopt the usual practice of using the region of residence of the individual as the geographic unit for ED events data as described in Section 3, some may deem this practice a limitation in our method. While the subregions identified were not necessarily the same for each year, for 2004/2005 and 2005/2006 fiscal years the results of each analysis were similar. Potential clusters were identified in sparsely populated north eastern region and the central region around the capital city (Edmonton). These zones may be true clusters or may represent areas where distributions of important factors are not the same (e.g., age distributions) and are not adjusted for in our analysis based solely on counts. The traditional spatial scan [19] applied to case counts failed to identify a particular subregion as part of the most likely cluster in 2004/2005 and this highlights how the assumptions of the data distribution can effect the conclusions. The differences among methods were further explored with simulated data sets where we showed that the traditional spatial scan based on case or event counts may not coincide with our compound Poisson spatial scan. It would be difficult to know *a priori* whether a traditional scan based on case counts or assumed independent event counts would provide similar results as a spatial scan assuming a compound Poisson data structure for the correlated events.

We recognize that the computational speed of the traditional spatial scan is superior, in that closed form expressions are available for only a few compound Poisson models and the denominator of the likelihood ratio test only has to be computed once, however, this advantage may be unimportant if the underlying model is not appropriate. Improvements could be made for computation. A few distributions of the compounding distribution (example, $Q(x;p) = -p^x/x\ln(1-p)$ for $x \geq 1$, and where $0 < p < 1$ would give $\Pr(U = u)$ a negative binomial distribution) are known to yield a closed form for the distribution of a compound Poisson random variable that will speed up the computation when evaluating the likelihoods instead of relying on the Panjer recursive formula [18]. In light of a closed form for the distribution of a compound Poisson random variable, and other possible choices of the compounding distribution, improvements in computation and examining performance represent further work for us to pursue. Without choosing a special compounding distribution to yield a closed form for $\Pr(U = u)$, computational speeds based on the Poisson or Bernoulli model always dominates.

As mentioned by many authors, cluster detection results depend on the value β which should be specified before the analysis. Our primary interest in this paper is discrete count data, however, one could easily extend our method to applications that deal with continuous data. As discussed in [5] and its references, the spatial scan statistic can be used for temporal data and be extended directly to a space-time setting for either retrospective or prospective analysis. We only looked at separate yearly analyses but further work can include formally extending our approach to spatio-temporal cluster detection. On the other hand, our current spatial scan test are developed to detect the change of case occurrence rates and parameter shift in the compound distribution $Q(.)$. In future development, we can consider testing for change in the form of $Q(.)$ under the alternative hypothesis.

Our approach allows for the detection of a most likely zone of disease-related (correlated) events in a geographic area based on a spatial scan and compound Poisson data assumptions. This approach will be useful for organizations, such as health administrators, who wish to identify geographic areas with higher numbers of disease-related events than expected. Upon identification, additional epidemiological investigations can be undertaken to determine a true cluster exists and if any policy interventions can be undertaken to reduce disease-related events.

Acknowledgement

The authors would like to thank the two anonymous reviewers for their comments that helped to strengthen some areas of this paper. The work was funded by an operating grant from the Canadian Institutes of Health Research. Rhonda J. Rosychuk is salary supported by Alberta Innovates - Health Solutions (AI-HS) as a Health Scholar. The authors thank Dr. Amanda Newton at the Department of Pediatrics at the University of Alberta for facilitating data use and for insightful comments.

This study is based in part on data provided by Alberta Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta. Neither the Government nor Alberta Health express any opinion in relation to this study.

References

1. Besag J, Newell J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A* 1991; **154**:143–155.
2. Kulldorff M, Nagarwalla N. Spatial disease clusters: Detection and inference. *Statistics in Medicine* 1995; **14**:799–810. DOI: 10.1002/sim.4780140809
3. Le ND, Petkau AJ, Rosychuk R. Surveillance of clustering near point sources. *Statistics in Medicine* 1996; **15**:727–740. DOI: 10.1002/(SICI)1097-0258(19960415)15:7/9<727::AID-SIM244>3.0.CO;2-X
4. Rosychuk RJ, Huston C, Prasad NGN. Spatial event cluster detection using a compound Poisson distribution. *Biometrics* 2006; **62**:465–470. DOI: 10.1111/j.1541-0420.2005.00503.x
5. Jung I, Kulldorff M, Klassen AC. A spatial scan statistic for ordinal data. *Statistics in Medicine* 2007; **26**(7):1594–1607. DOI: 10.1002/sim.2607
6. Rosychuk RJ, Stuber JL. An exact test to detect geographic aggregations of events. *International Journal of Health Geographics* 2010; **9**:1–14. DOI: 10.1186/1476-072X-9-28
7. Haase P. Spatial pattern analysis in ecology based on Ripley's K-function: introduction and methods of edge correction. *International Association of Vegetation Science* 1995; **6**(4):575–582. DOI: 10.2307/3236356
8. Nelson TA, Boots B. Detecting spatial hot spots in landscape ecology. *Ecography* 2008; **31**(5):556–566. DOI: 10.1111/j.0906-7590.2008.05548.x
9. Kumar MV, Chandrasekar C. Spatial clustering simulation on analysis of spatial-temporal crime hotspot for predicting crime activities. *International Journal of Computer Applications* 2011; **35**(3):36–43.
10. Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Satistical Society, Series A* 1991; **154**:421–441.
11. Lawson A, Biggeri A, Bohning D, Lesaffre E, Viel JF, Bertollini R. *Disease mapping and risk assessment for public health*. John Wiley & Sons, Chichester, UK, 1999.
12. Kulldorff M, Tango T, Park PJ. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* 2003; **42**(4):665–684. DOI: 10.1016/S0167-9473(02)00160-3
13. Kulldorff M. Tests of spatial randomness adjusted for an inhomogeneity. *Journal of the American Statistical Association* 2006; **101**(475):1289–1305. DOI:10.1198/016214506000000618
14. Openshaw S, Charlton M, Craft AW, Birch JM. Investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet* 1988; **331**(8580):272–273. DOI:10.1016/S0140-6736(88)90352-2
15. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 1990; **132**(1 Suppl):136–143.
16. Duczmal L, Assunção R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis* 2004; **45**:269–286. DOI: 10.1016/S0167-9473(02)00302-X
17. Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine* 1995; **14**(21-22):2323–2334. DOI: 10.1002/sim.4780142105
18. Panjer HH. Recursive evaluation of a family of compound distributions. *Astin Bulletin* 1981; **12**:22–26.
19. Kulldorff M. A spatial scan statistic. *Communications in statistics - theory and methods* 1997; **26**(6): 1481–1496. DOI: 10.1080/03610929708831995
20. **Matlab** version 7.14.0 (2012a). Natick, Massachusetts: The MathWorks Inc., 2012.
21. Newton AS, Rosychuk RJ, Ali S, Cawthorpe D, Curran J, Dong K, Slomp M, Urichuk L. *The Emergency Department Compass: Children's Mental Health. Pediatric mental health emergencies in Alberta, Canada: Emergency department visits by children and youth aged 0 to 17 years, 2002-2008*. Edmonton, AB, 2011. Retrieved December 24, 2012 from the World Wide Web: http://www.EDCompass.net
22. Kulldorff M and Information Management Services, Inc. **SaTScan™** v9.1.1: Software for the spatial and space-time scan statistics. http://www.satscan.org/, 2011.