STATISTICS IN MEDICINE, VOL. 14, 799–810 (1995)

SPATIAL DISEASE CLUSTERS: DETECTION AND INFERENCE

MARTIN KULLDORFF

*Department of Statistics, Uppsala University, Box 513, 751 20 Uppsala, Sweden, and Biometry and Field Studies Branch, National Institute of Neurological Disorders and Stroke, NIH, Federal Building 7C16, 7550 Wisconsin Avenue, Bethesda, MD 20892, U.S.A.**

AND

NEVILLE NAGARWALLA

Department of Dermatology, J-600, Boston University School of Medicine, 80 East Concord Street, Boston, MA 02118, U.S.A. and Bio Statistech, 15 Winthrop Road, Brookline, MA 02146, U.S.A.†

SUMMARY

We present a new method of detection and inference for spatial clusters of a disease. To avoid *ad hoc* procedures to test for clustering, we have a clearly defined alternative hypothesis and our test statistic is based on the likelihood ratio. The proposed test can detect clusters of any size, located anywhere in the study region. It is not restricted to clusters that conform to predefined administrative or political borders. The test can be used for spatially aggregated data as well as when exact geographic co-ordinates are known for each individual. We illustrate the method on a data set describing the occurrence of leukaemia in Upstate New York.

1. INTRODUCTION

The statistics of disease clustering is of interest to epidemiologists and has been studied for many decades. Such studies are useful to detect and monitor potential public health hazards. A review of several existing methods to detect spatial clustering of disease appears in Marshall[1] and in Hills and Alexander.[2] For more recent developments see Jacquez.[3]

The epidemiologist is typically interested in clusters of disease cases only after having adjusted for spatial variations in the density of the background population itself. Thus, on a map representing the cases as a spatial point pattern, an apparent disease cluster in a particular area could be misleading because it may be explained simply by a clustering of the population itself in that area. In this paper, we present a method that detects the location of possible disease clusters in a population with inhomogeneous spatial density, and simultaneously uses methods of inference to test for significance.

Upton and Fingleton[4] have pointed out two major approaches used for the analysis of spatial point patterns in general. Both have been applied to disease clustering. One approach uses a test statistic based on measuring distances between the disease cases while the other is based on

* Present address: Biometry Branch, DCPC, National Cancer Institute, EPN 344, 6130 Executive Blvd, Bethesda, MD 20892-7354, U.S.A.

CCC 0277–6715/95/080799–12

Received December 1993
Revised May 1994

studying the variability of case counts in certain subsets of the study region, often called quadrats. The former approach broadly defines the so-called distance methods, of which Whittemore *et al.*[5] is one example. Methods that rely on the latter approach are called quadrat-methods, an example of which appears in Choynowski.[6]

For the practitioner who intends to use a particular method, whether distance-based or quadrat-based, it is important to know exactly what the method can detect, partly because the term 'clustering' has several different interpretations. Most of the tests proposed so far have been *tests for overall clustering.* These do not have the ability to detect the location of clusters, but are geared toward answering the question of whether the phenomenon of clustering occurs in the data. Examples appear in Moran.[7] Whittemore *et al.*,[5] Cuzick and Edwards,[8] and Diggle and Chetwynd.[9] These are useful in applications where the location of clusters is not of interest, as for example, in an investigation of whether or not a disease is infectious.

In other situations one is interested in the location of clusters as well as in answering questions pertaining to their significance. We would then use what Besag and Newell[10] refer to as *tests for the detection of clusters.* Two primarily descriptive methods of this kind are those of Openshaw *et al.*[11,12] and Besag and Newell.[10] Both methods graphically identify possible clusters by using a multitude of overlapping circles as quadrats. Openshaw *et al.*[11] look at case counts in overlapping circles of variable size and identify potential clusters among these, by conducting a separate significance test for each circle individually. This method does not lend itself easily to a single unified test of significance, because the clusters identified in this manner are correlated and a Bonferroni procedure to compensate for multiple testing would be quite conservative. The test of Besag and Newell[10] uses overlapping circles to identify clusters in a slightly different way. In addition, it is combined with a test for overall clustering and thus appears to have a more acceptable statistical basis. Turnbull *et al.*[13] have used overlapping circles to construct a test that not only detects clusters but also correctly addresses the multiple testing problem, albeit only for circles with a pre-determined population size. The test we have developed in this paper generalizes the test of Turnbull *et al.*[13]

For yet another type of application we would use what Besag and Newell[10] call a *focused test.* One can use such tests when the study region contains some putative health hazard, such as a coal plant, and we suspect a cluster of say lung cancer around it. Examples appear in Stone,[14] Schulman *et al.*,[15] Diggle,[16] and Waller *et al.*[17] In the discussion in Section 5, we briefly mention how to adapt our method to do a focused test.

In Section 2, we describe the methods of Openshaw *et al.*[11,12] and Turnbull *et al.*,[13] since they relate closely to our method. Section 3 contains a precise statement of the null and alternative hypotheses and a description of the proposed likelihood ratio test for our method. The test we propose addresses several important problems. In particular:

1. We directly address the problem of inference for detected clusters.
2. We do not restrict ourselves to searching for clusters of a prespecified size.
3. The test is based on the likelihood ratio rather than an *ad hoc* test statistic.
4. We clearly define an alternative hypothesis so that the user of the test can decide whether the test is appropriate for the particular type of cluster detection problem at hand.
5. The method gives us a unique test statistic that makes it unnecessary to perform a separate test for each possible cluster location or each possible cluster size.
6. The test applies to aggregated as well as non-aggregated data.

In Section 4 we illustrate the method on a data set describing the incidence of leukaemia among the residents of Upstate New York. We end with a discussion in Section 5.

2. TESTS FOR THE DETECTION OF CLUSTERS

An early example of a quadrat-based test for the detection of disease clusters is the test proposed by Choynowski[6] which he applied to data on the distribution of brain-tumours in the Rzeszow province in Poland. As quadrats he simply takes the 17 different counties within the study region. He tests each quadrat individually to determine whether the number of cases in it is significantly high, at some level α. Testing each quadrat separately introduces the problem of multiple testing, but one can adjust for this by using a Bonferroni type procedure that would not be overly conservative. A more serious problem is our inability to detect clusters unless their boundaries coincide at least roughly with the county borders.

To overcome the above limitation, Openshaw *et al.*[11,12] developed a graphical method called the *geographical analysis machine* (GAM) that uses multiple overlapping circles of variable size as quadrats. One lays out a fine regular lattice of I points to cover the study region. The distance between adjacent lattice points is taken to be quite small. Then, one generates circular *zones*, centered at each lattice point i, ($i = 1, 2, \ldots, I$), and with a constant radius R that is typically 5 to 10 times the lattice spacing. Thus, there is considerable overlap between adjacent circles. For each circular zone, with centre i and radius R, the method requires determination of a critical value C_{iR}^*. This is taken to be the 99·8th percentile of the distribution of the random variable C_{iR}, the number of cases in the circular zone under the hypothesis that the cases distribute perfectly at random among the population. One considers circles where observed case count c_{iR} exceeds the critical value C_{iR}^* to have a significantly high incidence of the disease and then draws these 'significant' circles on the map. The procedure is repeated at three or four different values of R.

The technique used is hence identical to that of Choynowski[6] except that the quadrats overlap and are far more numerous. In data set similar to the one we consider in this paper, it is not unreasonable to have 100,000 or more circles. Then, although we no longer have to restrict our search to clusters that happen to coincide with some administrative boundary, as in Choynowski's method, now any Bonferroni type of procedure to adjust for multiple testing is futile due to the extremely large number of dependent tests performed. This method yields a very useful description of the data set with which one can identify several possible clusters.

Based on Openshaw *et al.*[11,12] Turnbull *et al.*,[13] have developed a test named the *cluster evaluation permutation procedure* (CEPP), that directly identifies the cluster responsible for causing rejection of the null hypothesis. The quadrats used in this method are once again overlapping circular zones. The circles centre at the geographic centroids of the K cells into which one has aggregated the data. Each circle, however, is constructed so as to have the same population size P, rather than the same radius. Here, it is useful to think of P as the 'population-radius' of the zones. Under the null hypothesis that the cases distribute randomly among the individuals of the population, the random variables C_{kP}, $k = 1, 2, \ldots K$, that represent the case counts in the various circular zones have identical probability distributions, but they are not independent. CEPP picks the zone with the highest incidence rate, or equivalently, the zone with the highest number of cases $M_P = \max\{C_{kP}: k = 1, 2, \ldots, K\}$ and then tests significance by using Monte Carlo simulation to sample from the null distribution of M_P. Thus CEPP uses the statistic M_P to test against a single composite alternative hypothesis whereas GAM would use each C_{kP} separately for multiple hypothesis testing. However, once P is fixed, the alternative hypothesis is that there is a 'cluster' among those circular zones of P persons that the method generates. Since there is no universal choice of P for all data sets, Turnbull *et al.*,[13] suggest that one should carry out their procedure at a few different representative values of P. This re-introduces multiple hypothesis testing, and, since the tests are highly correlated, a Bonferroni type adjustment is very conservative unless the number of different values used for P is very small.

The preceding remarks illustrate the crucial role played by the choice of zones in defining the alternative hypothesis, which is too often stated imprecisely as merely the opposite of complete spatial randomness. Further, the methods described above are limited by the difficulties associated with multiple testing. These are the lack of a unique test statistic and the consequent inability to assess quantitatively the overall significance of the results.

In the following section, we give a precise definition of our model which uses a composite alternative hypothesis in a single hypothesis test. The model builds upon the ideas contained in Openshaw *et al.*[11,12] and Turnbull *et al.*[13] We then present a test based on likelihood ratio.

3. A LIKELIHOOD RATIO TEST

Consider the study region partitioned into geographic sub-divisions called cells. For each cell, we have the co-ordinates of its geographical or population centroid, the number of individuals and the number of disease cases. The cell centroids form what Cressie[18] refers to as an irregular lattice. If the data are not aggregated at all, then each cell contains precisely one individual. We emphasize that we do not require any assumption about the population distribution within the cells. Let N be the total number of individuals in the population at risk and let C be the total number of cases. Throughout the analysis we condition on the total number of cases in the data set and hence we treat C as a known constant.

We can broadly classify the method of this paper as a quadrat method. Just as in two of the methods described in Section 2, we generate a number of 'circular' zones that we use as quadrats. To construct the circles, we have another lattice of I circle centers. This lattice could be regular as in Openshaw *et al.*[11] or identical to the irregular cell lattice as in Turnbull *et al.*[13] Unlike previous methods, for each centre point we let the radius of the circles vary continuously from zero upwards. Each of the infinite number of circles thus constructed defines a *zone*.

The zone defined by a circle consists of all individuals in those cells whose centroids lie inside the circle and each zone is uniquely identified by these individuals. Thus, although the number of circles is infinite, the number of zones will be finite. For unaggregated data the zones are perfectly circular, that is, the individuals in a zone are exactly those located within the defining circle. With the data aggregated into census districts, say, a zone may have irregular boundaries that depend on the size and shape of the several contiguous census districts it includes. Individuals actually outside the defining circle, but lying within cells whose centroids lie inside the circle, are included in the zone. Similarly, individuals actually inside the circle, but lying within cells whose centroids are outside the circle, are excluded. In any quadrat method, the alternative hypothesis is implicitly defined by the particular manner in which one constructs the zones or quadrats. This does not mean that the method can only handle the exact alternative defined. Rather, it gives an indication of the types of alternatives for which the test has good or bad power.

With an increasing radius, the circles will eventually include the entire study region. When a circle is so large as to include almost all of the study area it is inappropriate to talk about a cluster in that zone even if the incidence rate is considerably higher than outside of it. If anything, we could possibly view it as a kind of 'negative cluster' in those few areas that are still outside the circle. We do not wish to incorporate such negative clusters in the alternative and hence we need an upper bound on the radius of the circles to be considered. A natural rule of thumb that we advocate is 50 per cent of the total population. It is important to note that the choice should be made *a priori* and not by trial and error.

Denote by Z, the set of all circular zones generated in the manner described above. Let the (z, p, q) be a point in the parameter space where $p, q \in [0, 1]$ and z is a three-dimensional vector that

consists of the central co-ordinate and radius of a circle. We will interchangeably use z to denote both the vector itself and the zone it describes. In our model there is exactly one circular zone z, such that for all individuals within the zone, the probability of being a case is p, whereas for all individuals outside the zone, this probability is q. The alternative hypothesis is $H_1: z \in Z, p > q$. The null hypothesis is $H_0: p = q$. The latter signifies complete spatial randomness with each individual equally likely to be a case.

Let n_z denote the number of individuals in zone z, C_z the random variable denoting the number of cases in zone z and c_z the observed value of C_z in the data set. To derive the likelihood ratio test we first express the likelihood function which is

$$L(z, p, q) = p^{c_z}(1 - p)^{n_z - c_z} q^{C - c_z}(1 - q)^{(N - n_z) - (C - c_z)}. \tag{1}$$

Since the circular zones have different population sizes, we cannot merely take our test statistic as the maximum number of cases among all zones. It is not meaningful either, to take the maximum of the incidence rates among all circles since the variances of these quantities are unequal. Indeed, in many instances, the latter would lead us to pick the zone with the smallest number of individuals from among those zones that have at least one case. Instead, we use the likelihood ratio test statistic. The likelihood ratio is

$$\frac{\sup_{z \in Z, p > q} L(z, p, q)}{\sup_{p = q} L(z, p, q)} \quad (p, q \in [0, 1]). \tag{2}$$

The denominator in equation (2) reduces to

$$\sup_{p \in [0, 1]} p^C (1\text{-}p)^{N - C} = \frac{C^C (N - C)^{N - C}}{N^N} \stackrel{\text{def}}{=} L_0. \tag{3}$$

L_0 depends only on the total number of cases, not on their spatial distribution, and is a constant since we have conditioned on C. We can find the value of the numerator in two steps. First, for a fixed zone z, we maximize over all possible $0 \leqslant q \leqslant p \leqslant 1$. Let

$$L(z) = \sup_{p > q} p^{c_z}(1 - p)^{n_z - c_z} q^{C - c_z}(1 - q)^{(N - n_z) - (C - c_z)} \tag{4}$$

$$= \begin{cases} \left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{n_z - c_z}{n_z}\right)^{n_z - c_z} \left(\frac{C - c_z}{N - n_z}\right)^{C - c_z} \left(\frac{N - n_z - (C - c_z)}{N - n_z}\right)^{N - n_z - (C - c_z)}, & \text{if } \frac{c_z}{n_z} > \frac{C - c_z}{N - n_z} \\ \frac{C^C (N - C)^{N - C}}{N^N}, & \text{if } \frac{c_z}{n_z} \leqslant \frac{C - c_z}{N - n_z}. \end{cases} \tag{5}$$

As the most likely cluster, we pick the zone $\hat{z} \in Z$, for which the quantity defined by equation (5) is maximized. Formally, we choose $\hat{z}$ so that $L(\hat{z}) \geqslant L(z)$ for all $z \in Z$. This means that $\hat{z}$ is the maximum likelihood estimate of z. Identifying $\hat{z}$ is a necessary step for the likelihood ratio test, but it also has a purpose in itself if we have an interest in the descriptive aspects of the problem.

If we let $\mathbf{L}(z)$ denote the random variable obtained by replacing c_z with C_z in equation (5), then, combining equations (3) and (5), we can write the test statistic as

$$\lambda = \frac{\max_z \mathbf{L}(z)}{L_0}. \tag{6}$$

The distribution of λ depends on the underlying inhomogeneous population distribution, and, in general, it has no simple analytical form. If the total population is very small then it is possible to find the exact distribution by enumerating each of the possible outcomes, namely all the possible assignments of cases to the individuals of the population, and then computing the value of the test statistic for each outcome.

For large data sets, however, this is practically impossible, and thus we use the Monte Carlo method to sample from the exact distribution of λ. Note that we can do this easily since we have conditioned on the total number of cases C. The idea of significance testing based on the randomization distribution of a test statistic is due to Fisher.[19] The use of the Monte Carlo method for sampling from the randomization distribution to conduct a hypothesis test was suggested by Dwass.[20] It was first applied to spatial point patterns by Besag and Diggle.[21]

4. AN APPLICATION

The data set we analysed comes from Upstate New York, encompassing the counties of Broome, Cayuga, Chenango, Cortland, Madison, Onondaga, Tioga, and Tompkins. We have chosen this data set since it has previously been analysed in the literature by Turnbull *et al*,[13] using their own method as well as those of Whittemore *et al.*[5] and Openshaw *et al.*,[11] The same data have also been analysed by Waller *et al.*[17] and Waller and Turnbull[22] in the context of focused tests.

The data consists of 592 cases of leukaemia as represented in Figure 1. Since there is no information about the exact locations of individuals and cases, we have instead used the centroids of 790 census tracts and census block groups. Thus we are dealing with data aggregated into 790 cells. The data on the population counts and cell centroids are based on the 1980 US Census. The total population of the area is 1,057,673. Its distribution appears in Figure 2. Data on the leukaemia cases were obtained from the New York State Cancer Registry and cover the period from 1978 to 1982. There is some uncertainty as to the number of cases in each census area. For about 10 per cent of the cases, the location is known only within two or three neighbouring cells. Such cases were divided among the groups to which they may belong in proportion to the population in each group. This will tend to bias the conclusion away from clustering, but Turnbull *et al.*[13] have noted that this made little difference to their results. For the purely illustrative aspects of our methodology this uncertainty is irrelevant, but due to the manner in which it was resolved, some of the case counts have non-integer values.

For the set of zones Z upon which our alternative hypothesis depends, we use overlapping circles with center points at the centroids of the 790 census tracts. This follows Turnbull *et al.*[13] The radii of the circles vary continuously from zero, in which case we include only one cell, up to an upper limit, such that at most we include 20 per cent of the total population. This gives us an infinite number of circles, but, since the population is concentrated on 790 lattice points, we have a total of approximately $0{\cdot}2 \times 790^2 = 124{,}820$ distinct zones.

Our Monte Carlo study consisted of 999 replicates each of which involves choosing 592 individuals at random from the 1,057,673 individuals and labeling these as cases. For each replicate, we calculate the value of the test statistic λ defined in equation (6). We order the collection of 1000 values of λ coming from the 999 replicates and from the data itself with the highest value assigned rank 1. This means that we obtain a significant result at the 5 per cent level if the observed value of λ for the data is among the 50 highest of these 1000 values.

The observed value of the test statistic for the data is $\lambda = 472{,}976$. The most likely cluster is the zone $\hat{z} = \mathrm{A}$, shown in Figure 3. The rank of the observed λ value in the simulated null-distribution is 5 out of 1000. Thus, we have a significant result ($\alpha = 0{\cdot}05$) and we may attribute it to the presence of a cluster in zone A. Note that as long as the number of cases in zone A is 95·3, the

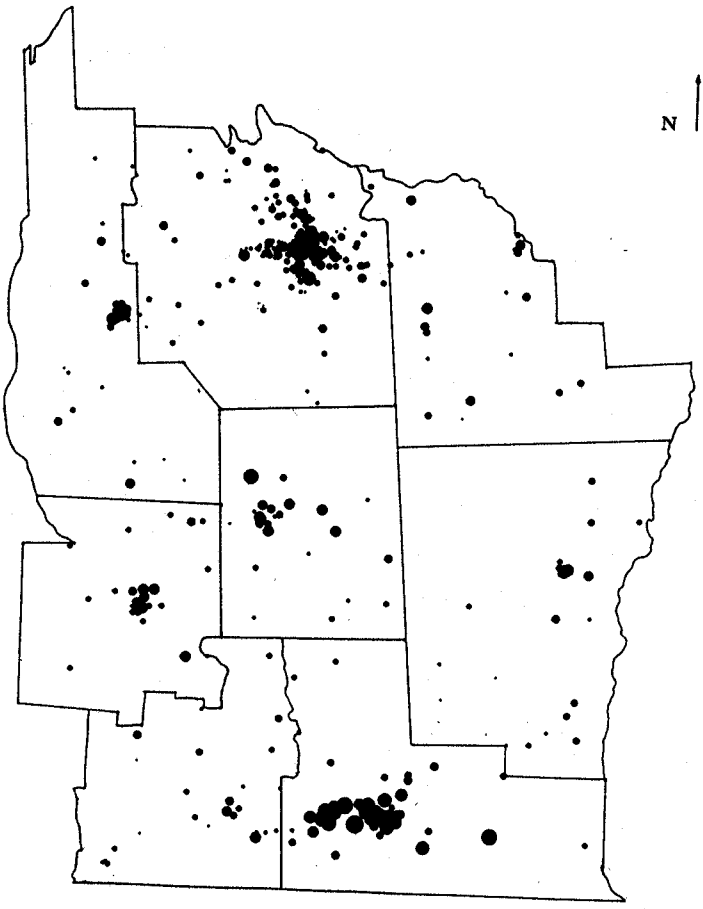

Figure 1. The 592 cases of leukaemia in Upstate New York

value of the test statistic can never be lower than its observed value of 472,976 even if the distribution of the cases outside zone A changes. This means that the cluster in zone A, by itself, ensures rejection of H_0.

It is important to realize that, even though zone A is the most likely cluster, it probably does not coincide exactly with the 'real' cluster. In any application there will be many zones almost identical to the most likely cluster for which the value of $L(z)$ is almost as high as $L(\hat{z})$. This is so because a change in the boundaries of a zone so as to include only a few more persons for instance, does not affect very much the value of $L(z)$. One should use the most likely cluster as an estimate for the position and radius of the real cluster in much the same manner as one would use a maximum likelihood estimate of an unknown parameter in a parametric hypothesis testing problem.

It is not of interest to report all zones with nearly equal values of $L(z)$. Table I lists the most likely cluster, A, along with four other non-overlapping clusters. For each of the zones B–D there is no other overlapping zone more likely to be a cluster. As is evident, not all of these zones have high ranking values of $L(z)$. The second zone, B, has a rank of 27 out of the 1000. If there had been no other more likely cluster, then we would have judged B significant ($\alpha = 0{\cdot}05$). We would,

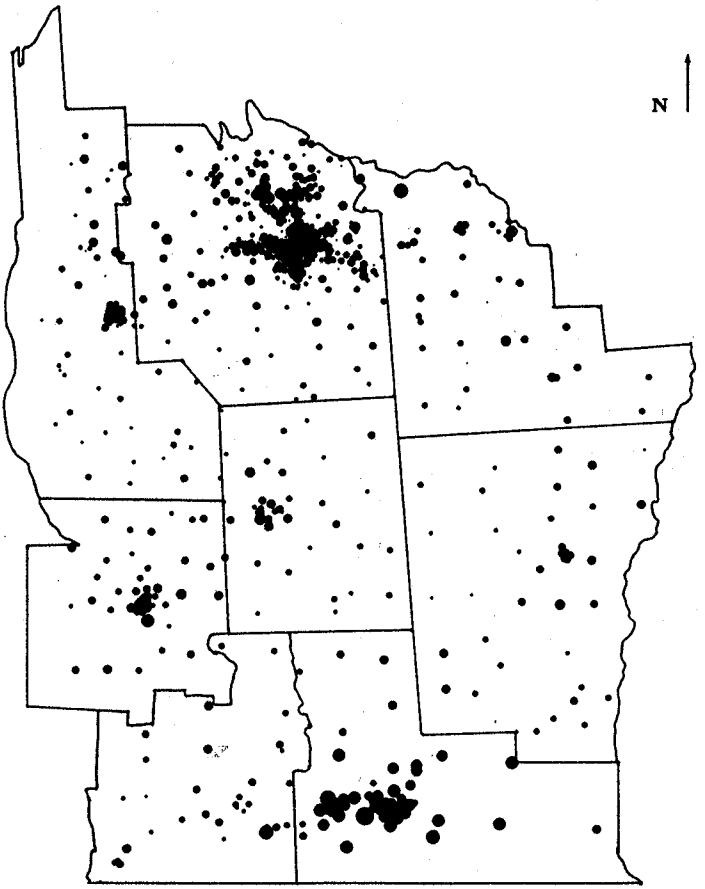

Figure 2. The population density in Upstate New York. The high density area in the north is Syracuse, and in the south Binghamton

however, then be assessing the value of $L(z)$ for the 'second most likely' non-overlapping cluster in the data set with reference to the distribution of $L(\hat{z})$ values that come from the most-likely clusters in the Monte Carlo replicates. That is, if we use the test for secondary clusters it is rather conservative. We could perhaps make our assessment with reference to the secondary clusters in the replicates but this would still be unsatisfactory, since it would not account for the size of the primary cluster in the data. The issue of secondary clusters is an interesting problem and deserves a fuller treatment.

Turnbull *et al.*[13] have applied several methods, including their own, on the same data set. Table II provides a comparative summary of their findings with the results of the likelihood ratio method. Despite the size of the problem and the large number of Monte Carlo replicates used, our implementation of the likelihood ratio method required only 2 hours of computing time on an IBM PC (PS/2 Model 90, XP486).

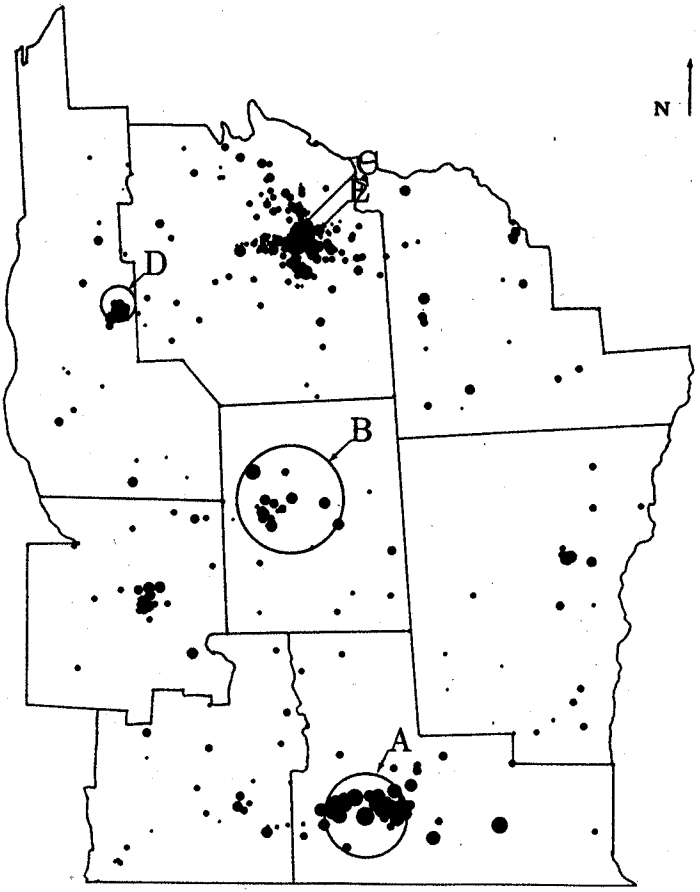

Figure 3. The most likely cluster 'A' and four other non-overlapping clusters on a map

5. DISCUSSION

In this paper we have given a general framework for the detection of spatial disease clusters and their evaluation using a likelihood ratio test. It was inspired by the introduction of overlapping circular zones as quadrats by Openshaw *et al.*[11] and the solution by Turnbull *et al.*[13] for circular zones with a fixed population size. We have emphasized the relationship between the manner of construction of the zones and the alternative hypothesis. Although we have described and implemented the method for circular zones of variable size, one can modify the likelihood ratio method for an alternative hypothesis that allows for zones of different shapes as well. The likelihood ratio test takes into account an inhomogeneous population density. It can also be modified to adjust for age-specific incidence rates. This would be necessary if the risk of the disease increases with age, say. These modifications will be described in forthcoming work. We conclude with the following observations:

1. When comparing the power of the likelihood ratio test with that of the method of Turnbull *et al.*,[13] one would except that (i) the latter has higher power if the actual cluster size is close

Table I. The most likely cluster A and four other non-overlapping clusters. The incidence rate for the population as a whole is 0·56

Zone z	Number of cases c_z	Population n_z	Incidence rate per 1000	Relative likelihood $L(z)/L_0$	Radius in km	Rank	County
A	95·3	99608	0·96	472976	6·3	5	Broome
B	43·2	36629	1·18	21088	10·2	27	Cortland
C	55·2	56806	0·97	1911	2·9	174	Onondaga
D	26·4	23682	1·11	187	2·8	781	Cayuga
E	3·4	793	4·29	51	0	996	Onondaga

Table II. Comparison of four methods on the leukaemia data set. Cluster size is given in population radius unless otherwise noted

Method	Cluster size in alternative hypothesis	Approximate cluster location	Significant ($\alpha = 0{\cdot}05$)
Openshaw *et al.*[11]	1,2,4 km	A B C D	n/a
Whittemore *et al.*[5]	n/a	n/a	no
Turnbull *et al.*[13]	2500	B	no
Turnbull *et al.*[13]	5000	B	no
Turnbull *et al.*[13]	10000	B	no
Turnbull *et al.*[13]	20000	B	yes
Likelihood ratio	$\leqslant 211535$	A B	yes
		C D E	no

Table III. Estimated power of the likelihood ratio test ($\hat{\beta}_l$) and of the method of Turnbull *et al.* ($\hat{\beta}_t$)

Cluster size	n_{cl}	100	200	400	700	1000	1400	2000	4000
Relative risk	rr	3·0	2·5	2·0	1·7	1·6	1·5	1·4	1·35
	$\hat{\beta}_l$	0·91	0·96	0·93	0·94	0·96	0·93	0·90	0·88
	$\hat{\beta}_t$	0·40	0·66	0·83	0·92	0·98	0·91	0·76	0·62

to the population-radius chosen in that method, and (ii) the likelihood ratio test has higher power for cluster sizes somewhat smaller or larger than this population-radius. A very simple power study presented in Table III confirms this. On a square, we selected randomly the locations of 100 cells. We assigned each cell a population of 100 to make a total of 10000 individuals. We placed another square with variable population size, n_{cl}, in the center to constitute the true cluster. We then randomly assigned 1000 cases among the population in such a way that individuals within the true cluster had a relative risk that was rr times higher than those outside. We set the circle size (population radius) at 1000 with the method of Turnbull *et al.*, and we used an upper bound of 5000 for our test. To obtain the power

estimates, we took 49,999 replicas from the null distribution and 5000 from each of the alternatives. Note that both methods perform well even though the real cluster is not circular.

2. There is a long tradition in epidemiology of publishing disease atlases, with incidence rates represented as different colors on a map. If one always complemented a disease atlas with an inference test for the detection of clusters, then public health officials could better prioritize the regions within which to conduct thorough investigations, with minimization of the time taken to detect genuine abnormalities. Once such a system is in place, one would perhaps like to have a sequential procedure for continuous monitoring.
3. The purpose of the new test for the detection of clusters, which we have presented here, is not limited to generating etiological studies. In many data sets we would find no significant cluster, but this can still be a very valuable finding. As Rothman[23] and many others have pointed out, vast resources are spent on the investigation of all possible alarms, often in vain, since many of these are plausibly explained as random fluctuations in the incidence rates. We do not imply that clusters that turn out non-significant with our method should never undergo investigation, but it could reduce the controversies that often occur with the reporting of potential clusters. One could then swiftly move resources to other more important tasks.
4. We wish to express a word of caution. The observed significance resulting from a particular cluster depends on the size of the area under study and it is not meaningful to attribute significance to a cluster without reference to the study region.
5. We mentioned in Section 3 that the centre points of the circular zones need not coincide with the locations of cells. If we were to pick only one centre point that coincides with the source of a possible health hazard, such as a coal plant or dump site, then we would have a focused test. In this case all zones have the same centre while the radius would still vary continuously. How such a focused test compares with existing methods merits further investigation.

ACKNOWLEDGEMENTS

We thank Dr. Phil Nasca, Director, Bureau of Cancer Epidemiology, New York State Department of Health, for providing the data set used in Section 4. We thank Bruce Turnbull, Lance Waller and Geoffrey Jacquez for valuable comments and discussions. This research was partially funded by the Swedish Council for Research in the Humanities and Social Sciences.

REFERENCES

1. Marshal R. C. 'A review of the statistical analysis of spatial patterns of disease', *Journal of the Royal Statistical Society, Series A*, **154**, 421–441 (1991).
2. Hills, M. and Alexander, F. 'Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: a review', *Journal of the Royal Statistical Society, Series A*, **152**, 307–325 (1989).
3. Jacquez, G. M. (ed) 'Proceedings of the Workshop on Statistics and Computing in Disease Clustering, Port Washington, New York, 1992', *Statistics in Medicine*, **12**, 1751–1968 (1993).
4. Upton, G. and Fingleton, B. *Spatial Data Analysis by Example, Volume* 1: *Point Pattern and Quantitative Data*, Wiley, New York, 1985.
5. Whittemore, A. S., Friend, N., Brown J., B.W. and Holly, E. A. 'A test to detect clusters of disease', *Biometrika*, **74**, 631–635 (1987).
6. Choynowski, M. 'Maps based on probabilities', *Journal of the American Statistical Association*, **54**, 385–388 (1959).
7. Moran P. A. P. 'The interpretation of statistical maps', *Journal of the Royal Statistical Society, Series B*, **10**, 243–251 (1948).

8. Cuzick J. and Edwards, R. 'Spatial clustering for inhomogeneous populations', *Journal of the Royal Statistical Society, Series B*, 52, 73–104 (1990).
9. Diggle, P. J and Chetwynd, A. D. 'Second-order analysis of spatial clustering for inhomogeneous population', *Biometrics*, **47**, 1155–1163 (1991).
10. Besag, J. and Newell, J. 'The detection of clusters in rare diseases', *Journal of the Royal Statistical Society, Series A*, **154**, 143–155 (1991).
11. Openshaw, S., Charlton, M., Wymer, C. and Craft, A. W. 'A mark 1 analysis machine for the automated analysis of point data sets', *International Journal of Geographical Information Systems*, **1**, 335–358 (1987).
12. Openshaw, S., Craft, A. W., Charlton, M, and Birch, J. M. 'Investigation of leukemia clusters by use of a geographical analysis machine', *Lancet*, **1**, 272–273 (1988).
13. Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L. and Clark, L. C. 'Monitoring for clusters of disease: application to leukemia incidence in upstate New York', *American Journal of Epidemiology*, **132**, S136–S143 (1990).
14. Stone, R. A. 'Investigations of excess environmental risk around putative sources: statistical problems and a proposed test', *Statistics in Medicine*, **7**, 649–660 (1988).
15. Schulman, J., Selvin, S. and Merrill, D. W. 'Density equalized map projections: a method for analysing clustering around a fixed point', *Statistics in Medicine*, **7**, 491–506 (1988).
16. Diggle, P. J. 'A point process modeling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point', *Journal of the Royal Statistical Society, Series A*, **153**, 349–362 (1990).
17. Waller, L. A., Turnbull, B. W., Clark, L. C. and Nasca, P. 'Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence and tce-contaminated dumpsites in upstate New York', *Environmetrics*, **3**, 281–300 (1992).
18. Cressie, N. *Statistics for Spatial Data*, Wiley, New York, 1991.
19. Fisher, R. A. *The Design of Experiments*, Oliver & Boyd, Edinburgh, 1935.
20. Dwass, M. 'Modified randomization tests for nonparametric hypotheses', *Annals of Mathematical Statistics*, **28**, 181–187 (1957).
21. Besag, J. and Diggle, P. J. 'Simple monte carlo tests for spatial pattern', *Applied Statistics*, **26**, 327–333 (1977).
22. Waller, L. A. and Turnbull, B. W. 'The effect of scale on tests for disease clustering', *Statistics in Medicine*, **12**, 1869–1884 (1993).
23. Rothman, K. J. 'A sobering start for the cluster busters' conference', *American Journal of Epidemiology*, **132**, S6–S13 (1990).