STATISTICS IN MEDICINE
Statist. Med. **18**, 1501–1516 (1999)

A COMPARISON OF SOME SIMPLE METHODS TO IDENTIFY GEOGRAPHICAL AREAS WITH EXCESS INCIDENCE OF A RARE DISEASE SUCH AS CHILDHOOD LEUKAEMIA

NAOMI R. WRAY[1], FREDA E. ALEXANDER[1]*, COLIN R. MUIRHEAD[2], EERO PUKKALA[3], IRENE SCHMIDTMANN[4] AND CHARLES STILLER[5]

[1] *Department of Public Health Sciences, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, U.K.*
[2] *National Radiological Protection Board, Chilton, Didcot, Oxon, OX11 0RQ, U.K.*
[3] *Finnish Cancer Registry, Liisankatu 21B, FIN-00170 Helsinki, Finland*
[4] *Johannes Gutenberg-Universität Mainz Klinikum, Institut für Medizinische Statistik und Dokumentation, 55101 Mainz, Germany*
[5] *Childhood Cancer Research Group, University of Oxford, Department of Paediatrics, 57 Woodstock Road, Oxford, OX2 6HH, U.K.*

SUMMARY

Six statistics are compared in a simulation study for their ability to identify geographical areas with a known excess incidence of a rare disease. The statistics are the standardized incidence ratio, the empirical Bayes method of Clayton and Kaldor, Poisson probability, a statistic based on the 'Breslow *T*' test (BT) and two statistics based on the 'Potthoff–Whittinghill' test (PW) for extra-Poisson variance. Two alternative processes of clustering are simulated in which high-risk locations could be caused by environmental sources or could be sites of microepidemics of an infectious agent contributing to a rare disease such as childhood leukaemia. The simulation processes use two parameters (proportion of cases found in clusters and mean cluster size) which are varied to embrace a variety of situations. Real and artificial data sets of small area populations are considered. The most extreme of the artificial sets has all areas of equal population size. The other data sets use the small census areas (municipalities) in Finland since these have extremely heterogeneous population size distribution. Subset selection allows examination of this variability. Receiver operator curve methodology is used to compare the efficacy of the statistics in identifying the cluster areas; statistics are compared for the proportion of true high-risk areas identified in the top 1 per cent and 10 per cent of ranked areas. One of the PW statistics performed consistently well under all circumstances, although the results for the BT statistic were marginally better when only the top 1 per cent of ranked areas was considered. The standardized incidence ratio performed consistently worst. Copyright © 1999 National Radiological Protection Board.

* Correspondence to: F. E. Alexander, Department of Public Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, U.K. E-mail: f.e.alexander@ed.ac.uk

Contract/grant sponsor: European Union
Contract/grant number: PL93-1785* 26.02.1993

Contract/grant sponsor: Leukaemia Research Fund
Contract/grant sponsor: Kay Kendall Leukaemia Fund

CCC 0277–6715/99/121501–16$17.50

Received December 1997
Accepted August 1998

INTRODUCTION

'*Post hoc*' cluster reports of rare diseases, such as childhood leukaemia, generate considerable public concern but are not readily amenable to formal statistical analysis. Nevertheless, public health professionals are often required to assess the evidence for excess risk, if any, to the local populations by comparing the reported cluster area to other geographically similar areas. Case-control studies have been undertaken to investigate putative cluster areas (see references 1–5). However, the number of cases which constitute individual clusters is small and these case-control studies have not led to conclusive results.

In contrast, regular examination of disease incidence databases for potential high-risk areas allows formal statistical analysis of these areas and more informative results may emerge from the comparison of high-risk to control areas. Such an approach must be taken cautiously to avoid arousing inappropriate public health concern. It can, however, make public health professionals and epidemiologists pro-active in assessing causes of clusters and could prepare the former for situations of intense public concern and media interest as has so often arisen for reported clusters of childhood leukaemia. For example, in the EUROCLUS project,[6–8] a study investigating clustering of childhood leukaemia in Europe, 20–25 area were identified from each participating region as having an excess of disease incidence. Information from lifestyle and environmental questionnaires was collected for these areas and compared to that from matched control areas (Alexander *et al.* 'Demographic factors in small areas containing clusters of childhood leukaemia: results of the EUROCLUS study', submitted for publication).

In both the reactive and pro-active situations described above, methods are required for accurate identification of high-risk areas. Interest will always focus on the highest ranking (say top 10 per cent) areas rather than accurate ranking of all areas. In the past, areas have been ranked simply by the ratio of observed (O_i) to expected number (E_i) of cases, O_i/E_i for each area i, or by the Poisson probability of the observed number of cases (for example, reference 9). It has been widely accepted for some time[10] that neither O_i/E_i nor Poisson probabilities are suitable for ranking areas for underlying risk of disease when geographical areas have low expected numbers of cases. No formal comparison of these methods or alternatives has been conducted.

Clayton and Kaldor[11] considered as an alternative to O_i/E_i for representation on disease incidence maps, a posterior estimate of the underlying relative risk; this is essentially a smoothing determined by the size and precision of O_i/E_i. It is based on an empirical Bayes approach which provides a clear ranking statistic. We note that fully Bayesian methods are also used extensively for mapping[12] with a variety of summary statistics plotted; we have not considered these methods here.

The 'Potthoff–Whittinghill'[13] and 'Breslow' T-statistics[14,15] have been used in recent studies[7,16] to test for extra-Poisson variance. They are simple tests to apply and differ in the form of the extra-Poisson variance expected in the alternative hypotheses for which they are optimal. The Potthoff–Whittinghill test has been shown in a study of artificial data[17] to perform well when compared to more complex and computer intensive methods in its ability to identify the presence of disease clustering. The theoretical properties and power of the Breslow T-statistic have also been investigated in detail.[15,18] Unlike some of the more complex methods (for example, reference 19), the Breslow T and Potthoff–Whittinghill tests do not identify individual cluster areas as a by-product of the detection of the presence of clustering. However, suitable functions of the posterior estimates of the underlying multinomial probabilities may be useful for ranking areas.

In this paper, a simulation of study compares six simple methods for ranking areas by their ability to identify known cluster areas. Real and artificial geographical census areas are used and cases are generated by simulation with clusters allocated by two different models.

METHODS

Small areas

The four data sets differ in the variability of population-at-risk (size) of small areas. The first three are taken from real small census areas (municipalities) in Finland; annual population counts were available in age group (0–4, 5–9, 10–14 years) and sex classes for the period 1980–1989. The E_i were derived by applying age and sex specific rates to the population at risk while maintaining the equality of ΣE_i with O, the total observed number of cases diagnosed 1980–1989. Finland is one of the countries participating in EUROCLUS which has small areas (municipalities) that are very variable in population count (see Figure 1). Data set TOT contains all municipalities. MEDV is restricted to those with between 0·1 and 5·0 expected cases of childhood leukaemia in the 10-year period, and so there is only medium variability in size of small areas. Alexander *et al.*[6] argued that this range of expected cases was appropriate for the clustering tests used in the EUROCLUS project. LOWV is restricted to municipalities with between 0·5 and 2·0 expected cases of childhood leukaemia in the 10-year period, and has mean childhood population close to that of data set TOT (Table I). The ratio of the mean to median E_i is 2·1, 1·5 and 1·2 for data sets TOT, MEDV and LOWV, respectively. Figure 1 shows the percentage of areas and population excluded from data set TOT to make data sets MEDV and LOWV. In each data set, T is the total number of areas ($T = 455, 413, 167$ for TOT, MEDV and LOWV, respectively). For each of TOT, MEDV and LOWV, the total O is that actually observed 1980–1989. The fourth data set, EQU is completely artificial; it has the same number of areas and cases as data set TOT, but the areas have equal E_i of O/T.

Simulation processes (see below) are used to allocate the O cells to the small areas with the 'observed' number in the ith being O_i.

Statistics ranking areas by evidence of excess risk

Six statistics are considered for ranking areas.

1. SIR, standardized incidence ratio, SIR $= O_i/E_i$, the maximum likelihood estimator of relative risk for each individual area under the basic Poisson model, in which O_i has Poisson distribution with mean $\lambda_i E_i$, where λ_i is the relative risk in the ith area.
2. EB, empirical Bayes; as in SIR, O_i has Poisson distribution with mean $\lambda_i E_i$ but the λ_i are assumed to be sampled from an underlying gamma distribution with mean of 1.

$$\text{EB} = (O_i + v)/(E_i + \alpha) = \left(\frac{O_i}{E_i}\right)\frac{E_i}{E_i + \alpha} + \left(\frac{v}{\alpha}\right)\frac{\alpha}{E_i + \alpha}$$

where v and α are defined in Clayton and Kaldor[11] and are estimated iteratively using their equations 5 and 7. Specifically, in the $j + 1$th iteration

$$v_{j+1} = \bar{\theta}_j^2/\beta_j \text{ and } \alpha_{j+1} = \bar{\theta}_j/\beta_j$$

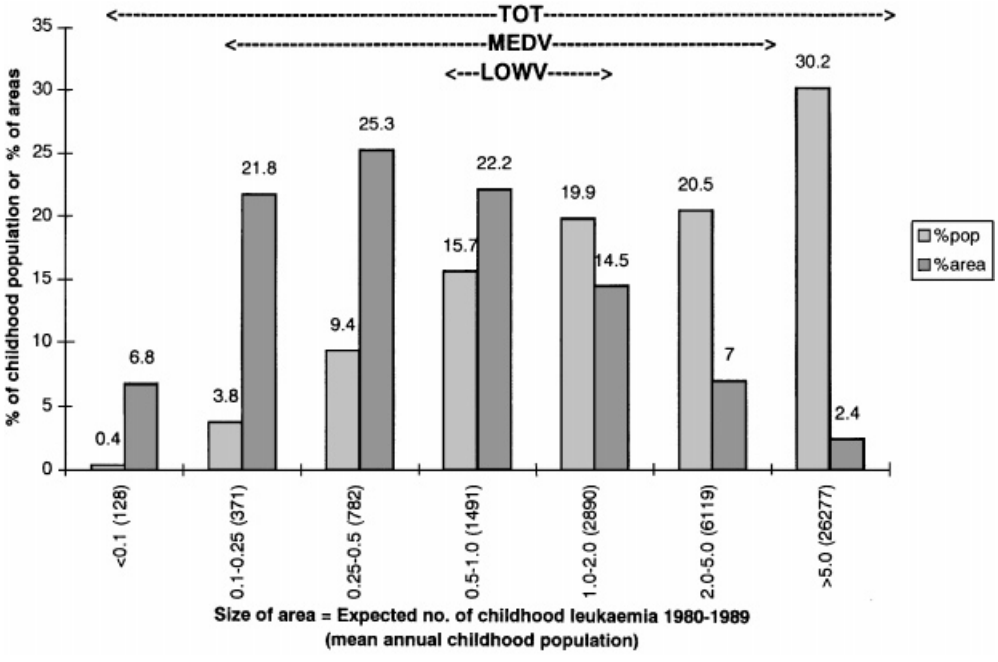

Figure 1. Distribution of size of small areas for data sets TOT, MEDV and LOWV

Table I. Summary statistics describing the small areas in the four data sets

	Data set			
	TOT	MEDV	LOWV	EQU
No areas (T)	455	413	167	455
No cases (O)	451	318	161	451
Childhood-population (annual) per area:				
Mean	2103	1607	2044	2103
Median	993	1041	1744	2103
Minimum	17	206	1006	2103
Maximum	73138	10453	4369	2103
Total population	956815	663814	341336	956815

where

$$\bar{\theta}_j = \frac{1}{T}\Sigma\theta_{ij}, \quad \beta_j = \frac{1}{T-1}\sum\left(1+\frac{\alpha_j}{E_i}\right)(\theta_{ij}-\bar{\theta}_j)^2$$

and

$$\theta_{ij} = \frac{O_i + v_j}{E_i + \alpha_j} \text{ with } v_0 = \alpha_0 = 0.$$

Convergence is reached when $\theta_{ij+1} - \theta_{ij} < 0{\cdot}001$ for all i. In practice, $v \cong \alpha$ and as $E_i \to 0$, EB $\to v/\alpha$ (if O_i is 0), but as $E_i \to \infty$, EB $\to$ SIR. When all E_i are equal $v = \alpha = 0$ and EB = SIR, but when the E_i are very variable the values of v and α increase.

3. Pois, Poisson probability of observing O_i or more cases in an area with expected number of cases E_i

$$\text{Pois} = 1 - \sum_{x=O}^{O_i - 1} \frac{e^{-E_i} E_i^x}{x!}$$

under the null hypothesis that O_i had distribution with mean E_i.

For the remaining three statistics we need to consider the form of the unconditional variance of O_i, $V(O_i)$. We shall take $V(O_i) = E(O_i) + X = E_i + X$.

4. BT, based on the Breslow T-statistic[14]

$$\text{BT} = (O_i - E_i)^2 - O_i \text{ for } O_i \geqslant E_i$$

and

$$\text{BT} = \text{an arbitrary working minimum for } O_i < E_i.$$

The score statistic for detecting extra-Poisson variation under a model such that X is proportional to ΣE_i^2 is related to $\sum[(O_i - E_i)^2 - O_i]$.[17,18] This is based on the unconditional distribution but with the true means replaced by $\{E_i\}$. The contribution of the ith area to this statistic can be large if $O_i < E_i$ (for example, if $O_i = 0$) and this is not informative. However, values based on $O_i > E_i$ can be considered for ranking purposes.

5. PW1, PW1 $= O_i(O_i - 1)/E_i^2$. The Potthoff–Whittinghill test for extra-Poisson variance is locally most powerful when X is proportional to E_i. For this test, each area contributes $O_i(O_i - 1)/E_i$ to the overall test statistic. This contribution is not a useful ranking statistic because areas with high values of O_i can be ranked highly even if $O_i < E_i$. A theoretical underpinning of the test takes $\{O_i\}$ multinomially distributed with parameters[20] $\{\lambda_i E_i, O\}$ and $\{\lambda_i\}$ a random sample from a gamma distribution with mean 1. Under this the posterior estimate of λ_i is $\{O_i(O_i - 1)\}^{0{\cdot}5}/E_i$. Thus ranking by PW1 is equivalent to ranking by the posterior estimates of the probabilities that an arbitrary case lies in the ith area. PW1 is the relative contribution of the ith area to the Potthoff–Whittinghill statistic. It is also suitable when the extra-Poisson variance is proportional to E_i^2 (see BT and process 2 below) since

$$E(\text{PW1}) = E\left(\frac{O_i(O_i - 1)}{E_i^2}\right) = \frac{1}{E_i^2}\left[V(O_i) + [E(O_i)]^2 - E(O_i)\right] = 1 + \frac{X}{E_i^2}.$$

6. PW2, PW2 $= O_i(O_i - 1)/E_i - E_i$ is another scale-free statistic based on the Potthoff–Whittinghill test focusing more on the absolute differences of O_i and E_i. PW2 is the contribution to the score statistic associated with testing for a multinomial distribution (conditional on the number of cases) against a Dirichlet-multinomial distribution, that is, unconditionally, the extra-Poisson variance, X is proportional to E_i as in process 1 below.[16] In this situation the expected value E(PW2) $= 1 + X/E_i$ is constant.

For all these statistics except Pois, the area most likely to contain a cluster will have the highest value. Figure 2 shows a simple comparison of the six ranking statistics; each figure shows the

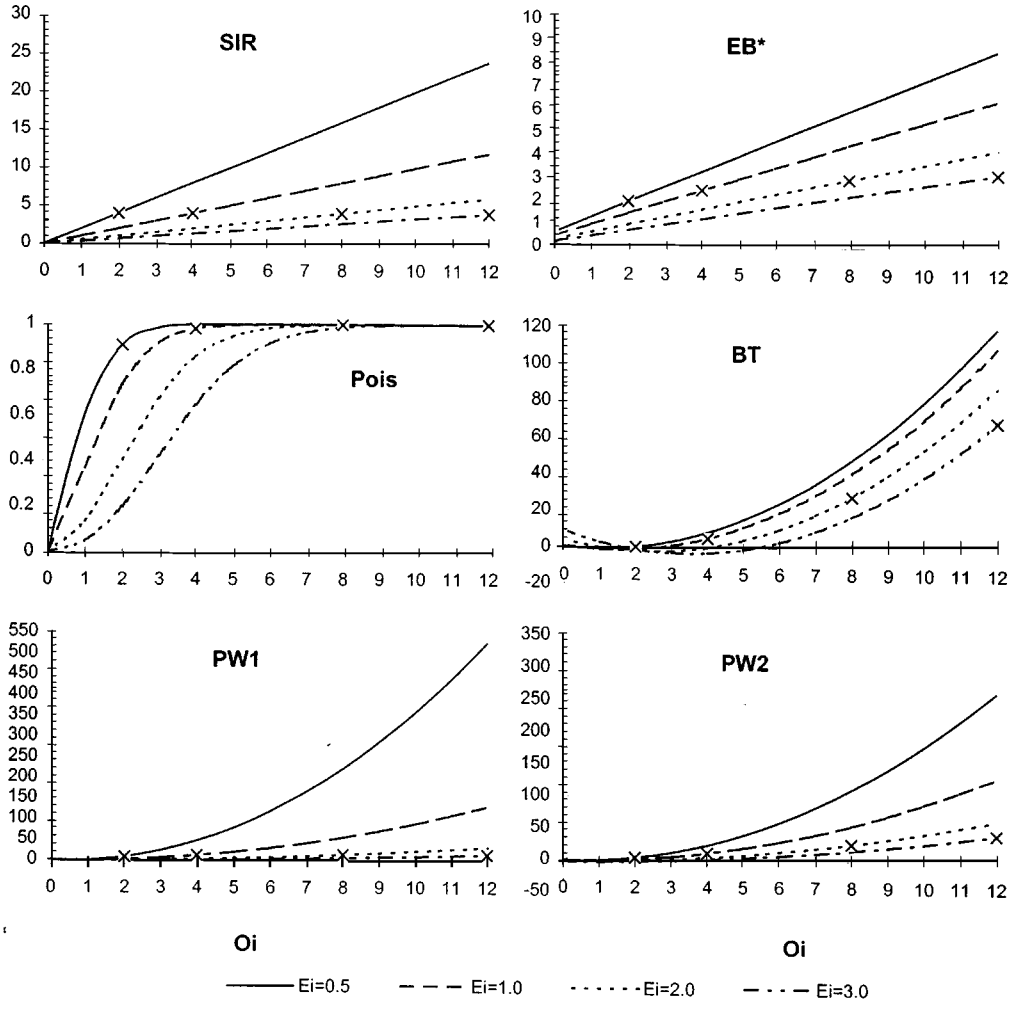

Figure 2. Relation between value of ranking statistic and observed number of cases for $E_i = 0{\cdot}5, 1{\cdot}0, 2{\cdot}0, 3{\cdot}0$, '$\times$' mark the points where $O_i/E_i = 4$: y-axis, value of each statistic; x-axis, O_i

relation between the value of the ranking statistic and O_i when E_i is 0·5, 1·0, 2·0, 3·0. Interest lies in the shape of the curves and the relative positioning of each E_i contour. The graph for EB uses $\nu = \alpha = 1$; the relative distance between the contours can be changed by altering the values of ν and α. The crosses mark the point on each contour line where $O_i/E_i = 4{\cdot}0$. For constant O_i/E_i, the value of BT increases with E_i while the values of SIR are independent of E_i; the other statistics are intermediate between these two extremes.

Methods to simulate clustering

The six ranking statistics are compared for situations of known clustering in which a simulation process allocates an 'observed' number of cases O_i to each area i. The models used to simulate

Table II. The alternative processes for simulation of clustering

Process number	Choice of high-risk locations	Poisson mean for number of cases/high-risk location in ith area*	(Unconditional) mean† of O_i $E(O_i)$	Extra-Poisson component of variance O_i
1	Randomly from population-at-risk‡	μ	E_i	$(\mu q)E_i$
2	Randomly from list of small areas§	$\mu E_i/\bar{E}$	E_i	$(\mu q/\bar{E})E_i^2$

* $\bar{E} = \Sigma E_i/T$, T is the total number of small areas;
† A proportion $(1 - q)$ of the total cases are allocated at random to the population at risk and the remainder by the clustering process listed. O_i, $E(O_i)$ refer to *all* cases in the ith area: O_i is the observed number of cases and E_i is the number expected under the Poisson variability ($\Sigma O_i = \Sigma E_i$);
‡ Each person at risk has an equal probability of selection; §Each small area has an equal probability of selection

clustering of a rare disease are Neyman centre-satellite processes;[21] they are the same as processes 1 and 2 presented by Alexander *et al.*[6] for the investigation of the power of cluster detection and use two parameters μ (mean cluster size) and q (proportion of cases allocated to clusters) in a two-stage process:

1. A number, h, of 'high-risk' locations are identified at random, where $h = qO/\mu$; h is an integer within a simulation replicate but the expected value is maintained over simulation replicates. A small area may be selected more than once to contain a high-risk location (and could therefore contain several high-risk locations). Areas not selected to contain high risk locations are considered to be 'standard' risk areas.
2. Each high-risk location generates a number of clustered cases in its own area with number of cases being sampled from a Poisson distribution; an area is still considered to be high risk even if the number of clustered cases generated is zero. In this event, data analyses will be incapable of detecting the area.

The remaining $(1 - q)O$ cases are allocated randomly to the population at risk. Two alternative processes for generation of clustering are considered (see Table II). Both processes maintain independence of population size and incidence rates ($E(O_i) = E_i$). Process 1 has variance $V(O_i) = E_i + cE_i$ (for which the test for extra-Poisson variance proposed by Potthoff and Whittinghill[13] is locally most powerful), whilst the variance generated by process 2 is $V(O_i) = E_i + cE_i^2$ (for which the Breslow T-test is locally most powerful). In process 2 the disease clusters in the geographical areas with highest E_i are expected to be bigger than in process 1.

A range of values for $q(0{\cdot}05, 0{\cdot}15)$ and $\mu(0{\cdot}5, 1{\cdot}0, 1{\cdot}5)$ are considered which are believed to be realistic for clustering of childhood leukaemia and specific combinations appropriate to the observed range of extra-Poisson variance at 2 per cent ($q = 0{\cdot}02$, $\mu = 1{\cdot}0$ or $q = 0{\cdot}04$, $\mu = 0{\cdot}5$) to 10 per cent ($q = 0{\cdot}10$, $\mu = 1{\cdot}0$ or $q = 0{\cdot}20$, $\mu = 0{\cdot}5$) observed in the EUROCLUS project.[7] Reported results are the average of 10,000 independent simulation replicates.

Method to compare ranking statistics

The method for comparing the efficiency of the six statistics in identifying cluster areas is based on the methodology of receiver operating characteristic (ROC) curves (see reference 22 for a review).

In its standard medical usage, the x-axis of the ROC curve is the 'false positive ratio' (or 1-specificity) and the y-axis is the 'true positive ratio' (or sensitivity). Here, the 'false positive ratio' is the proportion of standard-risk areas that are selected by the ranking process, and the 'true positive ratio' is the proportion of high-risk areas selected by the ranking process. An area may be high risk but have no cases generated in the clusters; such areas will be counted as false negatives unless sufficient random cases have been allocated to them.

For each statistic n ranking categories were created. Preliminary runs found working minimum and maximum values for each statistic. Three-quarters of the ranking categories were allocated equally to the first quartile of the range and one-quarter allocated equally to the remaining three quartiles. The number of small areas with statistic value falling in each ranking category was counted separately for areas with and without high-risk locations. (Ranking category rather than rank was used to ensure compatibility when averaging over simulation replicates.) The ROC curve has n points generated. For the jth point, the x co-ordinate is the number of standard-risk areas that have statistic values falling in the first j ranking categories expressed as a proportion of the total number of standard-risk areas. Similarly, the y co-ordinate is the number of high-risk areas that have statistic values falling in the first j ranking categories expressed as a proportion of the total number of high-risk areas.

A statistic which perfectly separates high- and standard-risk areas in ranking order has a unit square ROC curve (Figure 3(a)) and a statistic which ranks areas randomly is expected to have a unit diagonal ROC curve (Figure 3(b)). A statistic which ranks some high-risk areas highly and then ranks the remainder randomly with standard-risk areas (Figure 3(c)) is considered specific but not very sensitive, whereas a statistic which ranks some of the standard risk areas last but for which the remainder are ranked randomly interdispersed with the high-risk areas (Figure 3(d)) is considered sensitive but not very specific. In practice, the best that can be achieved will depend on the proportion of high-risk areas that have at least one case allocated and the main interest is to ensure that the areas ranked highly are truly high-risk areas (that is, specificity is important). Therefore, good ranking statistics will result in ROC curves of the type Figure 3(c) tending as much as possible to the shape of Figure 3(a).

Values of the (x, y) co-ordinates for each category were averaged over replicates and are used for graphical presentation of the curves. If the top ranking vT areas ($v = 0{\cdot}01, 0{\cdot}10$) are chosen, the proportion of these areas which are truly high-risk locations is $y_v h/vT$ (the positive predictive value) where y_v is found by interpolation to satisfy $x_v^*(T - h) + y_v^* h = vT$. The number of ranking categories is $n = 500$ and preliminary runs demonstrated that increasing n did not change the results to the accuracy to which they are reported.

As a partial verification of the simulation program, an extreme example was considered in which all small areas have equal E_i and all cases are allocated to clusters $q = 1{\cdot}0$. In this simple situation (process 1 = process 2, and the ranking of areas is identical for some statistics SIR = EB = Pois (and this corresponds to BT for areas with $O_i > 1{\cdot}5E_i$), PW1 = PW2) simulation results could be predicted from binomial and Poisson probabilities.

RESULTS

Figure 4 shows the ROC curves for the six ranking statistics and for the two clustering processes with $q = 0{\cdot}15$ and $\mu = 1{\cdot}0$ for data set TOT. For process 1 there are clear differences between the ranking statistics and their relative superiority depends on the proportion of all areas (that is, the distance from the origin) to be considered. For process 2, the methods perform more similarly and

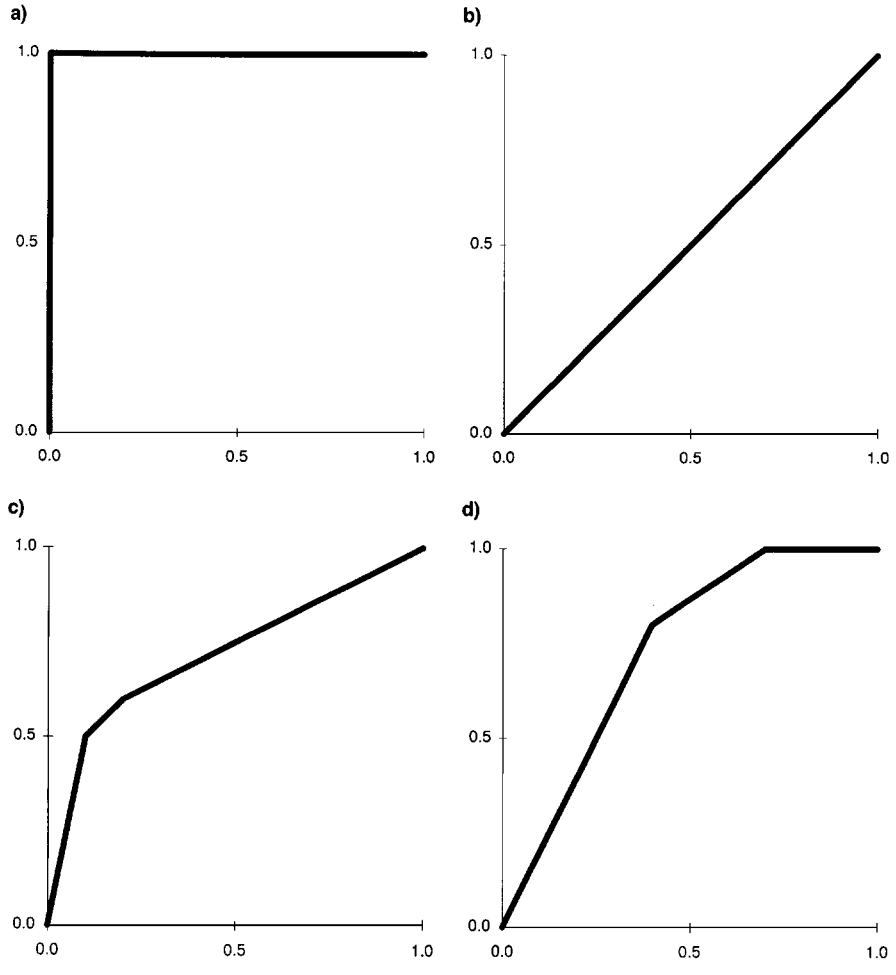

Figure 3. Possible shape of ROC curves: x-axis, proportion of all standard-risk areas included in the ranking = false positive ratio = 1-specificity; y-axis, proportion of all high-risk areas included in the ranking = true positive ratio = sensitivity

reflect the fact that the proportion of high-risk areas with zero cases is about 0·45 compared to only about 0·10 in process 1.

Three of the methods rank a group of areas equal last (any areas with zero cases in SIR, any areas with $O_i < E_i$ in BT and any areas with less than two cases in PW1); consequently, their curves have a 'turning point' with a straight line drawn from this point to (1, 1). For the other methods, areas are distinguished by their E_i and areas with low E_i and zero cases may be ranked higher than, for example, an area with high E_i and only a single case. SIR may rank highly areas with a very small E_i which, by chance, contain a single case; this explains the gentle rise of its ROC curve from the origin. In contrast, the ROC curve of EB discourages high ranking of areas with low E_i and single cases and so the ROC curve rises more steeply from the origin. Pois is better

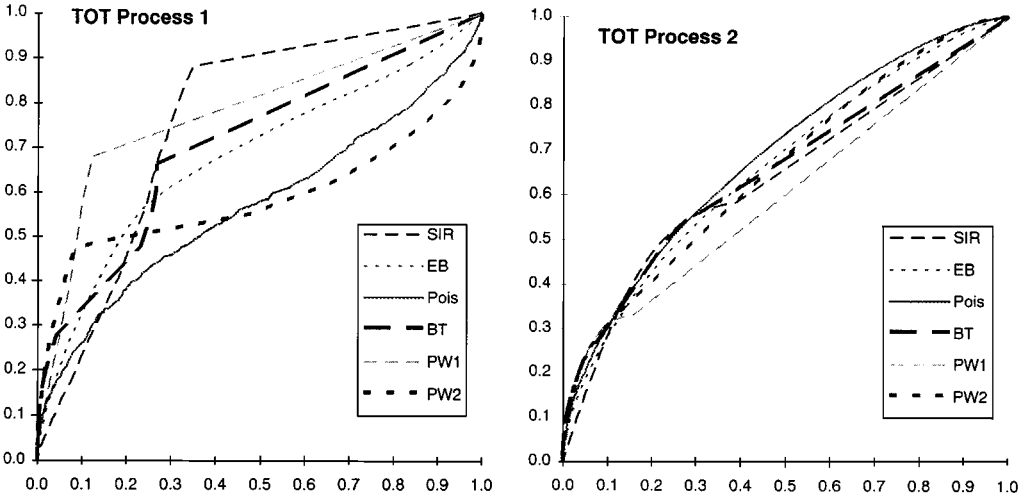

Figure 4. ROC curves for the six ranking statistics for processes 1 and 2, $q = 0{\cdot}15$ $\mu = 1{\cdot}0$ for data set TOT: x-axis, proportion of all standard-risk areas included in the ranking; y-axis, proportion of all high-risk areas included in the ranking

than SIR and EB in identifying high-risk areas if only the very first ranked areas (< 1 per cent) are considered, but for clustering process 1 it performs much worse than other methods if there is interest in ranking more areas. For both processes, PW2 and BT (and to a lesser extent PW1) rise steeply from the origin and so these methods are more likely to rank high risk areas in the first 10–15 per cent of areas. Interest of correct identification of high-risk areas is likely to be limited to only a proportion of areas, therefore subsequent tabulated results consider only the top 10 per cent and top 1 per cent of ranked areas.

The effect of variability in the (population-at-risk) size of small areas on the ROC curve is shown for process 1 in Figure 5. As the small areas become more uniform in size the difference in statistics in ranking areas is reduced, so that in data set EQU SIR = EB = Pois (= BT when $O_i > 1{\cdot}5E_i$) and PW1 = PW2 (equalities based on rankings of areas not values). However, there is still an important difference in efficacy of identifying high-risk areas between statistics for process 1 and between processes in the data set which has low variability in size of areas (LOWV). Results for ranking of 1 per cent and 10 per cent of areas are listed in Table III.

The effect of changing q and μ within a range appropriate to clustering of a rare disease is shown in Table IV for selection of 10 per cent and 1 per cent of areas. As q (the proportion of cases found in clusters) increases and as μ (the mean size of cluster) decreases, the proportion of areas selected which are high-risk areas increases, partly because total number of areas with clusters has increased. The relative performance of the ranking statistics is robust to the values of μ and q for the range of combinations considered.

For each country in the EUROCLUS project, 20–25 areas were selected using the PW2 criterion. Table V shows for Finland, the percentage of these areas expected to be true high-risk areas based on these combinations of q and μ which could result in the 2–10 per cent extra-Poisson variance observed in the EUROCLUS analysis.

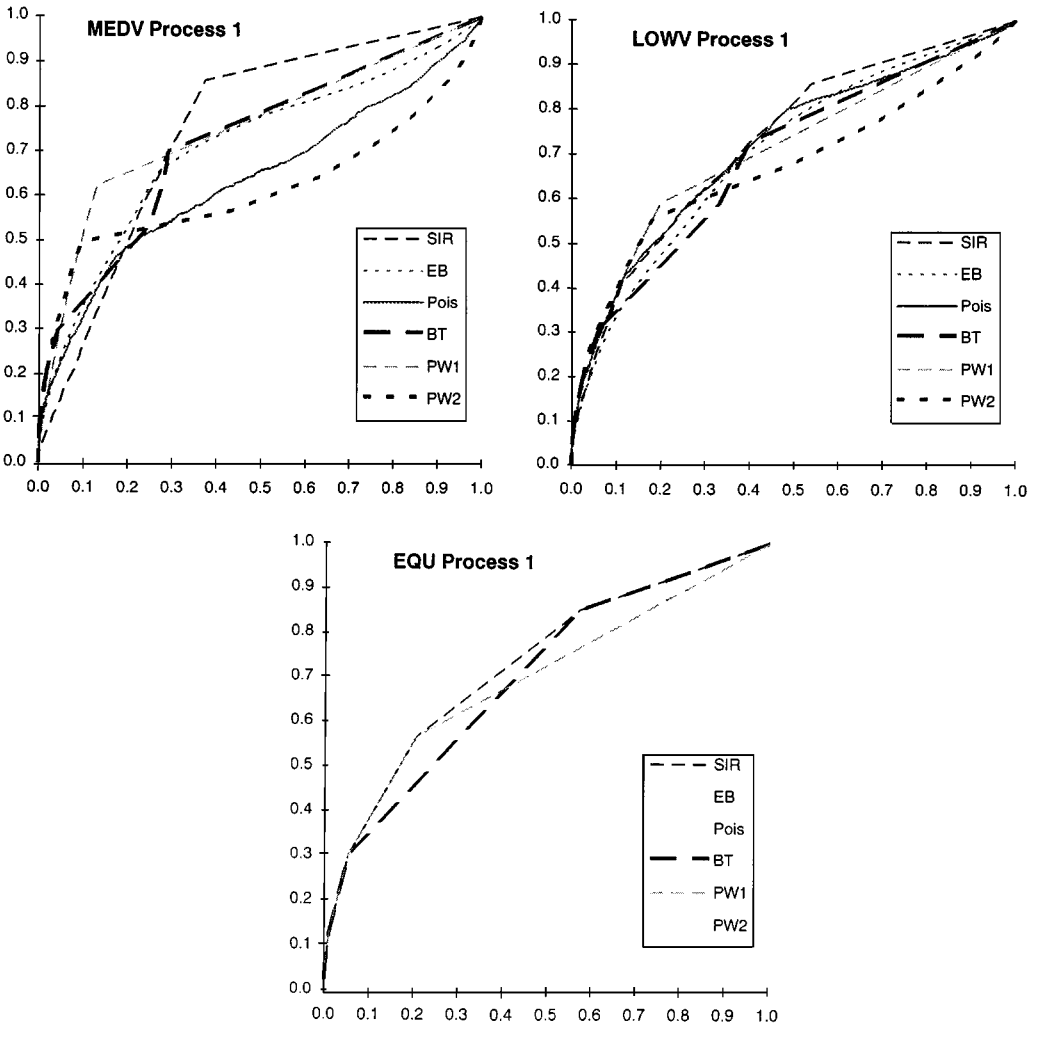

Figure 5. ROC curves for the six ranking statistics for data sets MEDV, LOWV and EQU for process 1, $q = 0{\cdot}5$ $\mu = 1{\cdot}0$: x-axis, proportion of all standard-risk areas included in the ranking; y-axis, proportion of all high-risk areas included in the ranking

DISCUSSION

Six ranking statistics have been compared in analyses of real and artificial geographical areas for two alternative processes of rare disease clustering. Process 1 is motivated by the population mixing ('virus') hypothesis of clustering of a rare disease such as childhood leukaemia in which the probability of a small area containing a high-risk location depends on the population at risk. Areas with high E_i may contain several high-risk locations, but the size of cluster generated from each high-risk location is independent of population at risk. In contrast, process 2 may represent

Table III. Percentage of areas ranked in the top 10 per cent and top 1 per cent that are truly high risk areas, $q = 0{\cdot}15$, $\mu = 1{\cdot}0$

Data set	T*	h†	Process	10% of areas selected						1% of areas selected					
				SIR	EB	Pois	BT	PW1	PW2	SIR	EB	Pois	BT	PW1	PW2
TOT	455	67·6	1	29	49	35	45	50	55	39	83	70	87	64	84
			2	34	40	38	39	41	40	36	79	68	81	54	79
MEDV	413	47·7	1	26	42	33	38	43	45	42	78	70	79	59	77
			2	35	40	39	39	40	40	42	83	76	83	58	82
LOWV	167	24·4	1	42	45	44	45	44	47	68	74	75	76	70	76
			2	77	79	79	79	79	79	91	96	95	96	93	96
EQU	455	67·6	1,2	47	47	47	46	47	47	78	78	78	78	78	78

* Total number of geographical areas;
† Expected number of high-risk locations (some areas may have more than one high-risk location)

Table IV. Effect of q and μ on the percentage of areas ranked in the top 10 per cent and top 1 per cent that are truly high risk areas for data set TOT

Process	q	μ	10% of areas selected						1% of areas selected					
			SIR	EB	Pois	BT	PW1	PW2	SIR	EB	Pois	BT	PW1	PW2
1	0·05	0·5	12	24	14	21	25	30	12	46	27	52	24	47
		1·0	9	17	11	15	18	21	13	43	30	48	26	46
		1·5	8	15	10	13	16	18	14	44	34	50	28	48
	0·15	0·5	39	61	44	58	63	69	40	83	66	89	63	83
		1·0	29	49	35	45	50	55	39	83	70	87	64	84
		1·5	26	42	31	40	44	49	41	86	75	89	67	87
2	0·05	0·5	16	17	18	19	19	19	17	33	28	38	23	37
		1·0	12	13	14	14	14	14	13	35	31	38	22	39
		1·5	11	11	12	12	12	12	12	36	35	39	24	40
	0·15	0·5	45	50	48	51	53	53	46	76	66	81	58	79
		1·0	34	40	38	39	41	40	36	79	68	81	54	79
		1·5	31	35	34	34	35	35	34	81	72	81	55	81

Table V. The percentage of areas ranked in the top 5 per cent (22·6 areas) that are truly high-risk areas for data set TOT, for statistic PW2 and for values of q and μ relevant to the amoung of extra-Poisson variance found in the EUROCLUS project

	2% extra-Poisson variance		10% extra-Poisson variance	
	$\mu = 1$ $q = 0{\cdot}02$	$\mu = 0{\cdot}5$ $q = 0{\cdot}04$	$\mu = 1{\cdot}0$ $q = 0{\cdot}10$	$\mu = 0{\cdot}5$ $q = 0{\cdot}20$
Process 1	12	8	50	85
Process 2	29	19	37	74

the fixed environmental hazard ('chimney') hypothesis of clustering of a rare disease: each geographical area, regardless of population at risk, has equal chance of having a high-risk location, but the size of cluster is generated from a distribution dependent on population-at-risk. For both processes the unconditional expectation of O_i is equal to E_i. Processes 1 and 2 generate the same number of high-risk locations, but process 2 will have more high-risk locations without cases present as well as some very large clusters in areas of high E_i. When all areas have equal E_i the processes are the same; in this situation the six ranking statistics are identical in their ranking of areas with two or more cases (and if $O_i > 1{\cdot}5E_i$). However, as the variability in E_i increases to levels found commonly in real census data, differences between the six ranking statistics become apparent. In particular, if process 1 is most likely to represent the true model of exposure, then the choice of ranking statistic is critical. In practice, it is important that the chosen statistic ranks highest areas that are truly high risk (that is, it must be specific) rather than trying to rank all high-risk areas before standard-risk areas (that is, sensitivity is less important). Accurate ranking of up to 10 per cent of areas is likely to be of interest (Table III). Over the range of situations considered in this study, PW2 has performed best for all data sets and both processes when interest is in accurate ranking of 10 per cent of all areas. EB and BT also perform well, but are able to achieve this by identifying cluster areas with only a single case. In contrast, PW2 must be identifying more high-risk areas with larger clusters, which are likely to be of more interest in practice. When interest is in accurate ranking of only the top 1 per cent of areas, PW2 again performs well, but BT performs marginally better and with so few areas selected, those chosen are likely to have high E_i. Pois matches the performance of PW2 when only 1 per cent of areas are selected in data set LOWV, but performs less well when area sizes are more variable and a higher proportion of areas selected. PW1 performs well when 10 per cent of areas are selected but is not as specific as PW2, EB and BT when only 1 per cent of areas are selected. SIR performs worst and should be avoided in all situations.

The results are appropriate to situations where processes 1 and 2 are likely to approximate the distributions of interest, especially as they relate to the high-risk areas. We note, in particular, that the EB method was derived for the situation in which the $\{\lambda_i\}$ were sampled from a gamma distribution with mean 1 ($\Gamma(1)$). We have examined the empirical distributions for $\{\lambda_i\}$ in process 1 (data not shown) and found the variability to be greater than predicted by $\Gamma(1)$. This is largely attributable to larger percentages of areas having $\lambda_i < 1$ in process 1; the conditional distributions of λ_i for $\lambda_i > 1$ for process 1 and $\Gamma(1)$ agree quite closely. Thus, it is likely that our results would apply if we had used $\Gamma(1)$ to generate $\{\lambda_i\}$ and hence $\{O_i\}$. However, we cannot exclude the possibility that EB would have performed optimally in this situation. The processes used to

generate the present clustering have been chosen as having biological and empirical rationale; they are probably more appropriate than $\Gamma(1)$ for the situations where the selection of a relatively small number of genuinely high-risk areas is important.

The ranking statistics and simulation processes used in this study reflect our interest in localized clustering *within* small areas; data and confidentiality rules common to all countries participating in the EUROCLUS project means that only *within* small area analysis was possible. Of course, biological or environmental causes of clustering of a rare disease are unlikely to respect artificial census boundaries, but this only serves to dilute the clustering that can be detected. In the EUROCLUS project we selected 20–25 areas from each participating region using PW2, which for Finland represents about 5 per cent of areas. Between 2–10 per cent of extra-Poisson variance was detected in the overall EUROCLUS analysis[7] which may mean that anything from 8–85 per cent of the areas chosen for further study may be truly high-risk areas if the spread of disease followed one of the processes considered here. Significant results from the comparison of the suspected cluster areas to matched control areas (Alexander *et al*., 'Demographic factors in small areas containing clusters of childhood leukaemia: results of the Euroclus study' and 'Population density and childhood leukaemia: results of the EUROCLUS study', in press (*Eur. J. Canc*.) and examination of temporal patterns in the cluster areas provides verification that the method is useful.

ACKNOWLEDGEMENTS

The co-ordination of this project if funded by the European Union under its BIOMED programme of Concerted Actions as project number PL93-1785*26.02.1993. Dr. Freda Alexander is also partially supported by the Leukaemia Research Fund and the Kay Kendall Leukaemia Fund. The Childhood Cancer Research Group is supported by the UK Departments of Health.

REFERENCES

1. Gardner, M., Snee, M., Hall, A., Powell, C., Downes, S. and Terrell, J. 'Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria', *British Medical Journal*, **300**, 423–434 (1990).
2. Urquhart, J., Black, R., Muirhead, M., Sharp, L., Maxwell, M., Eden, O. and Jones, D. 'Case-control study of leukaemia and non-Hodgkin's Lymphoma in children in Caithness near the Douneray nuclear installation', *British Medical Journal*, **302**, 687–692 (1991).
3. Roman, E., Beral, V., Carpenter, L., Watson, A., Barton, C., Ryder, H. and Aston, L. 'Childhood leukaemia in the West Berkshire and Basingstoke and North Hampshire District Health Authorities in relation to nuclear establishments in the vicinity', *British Medical Journal*, **294**, 597–602 (1987).
4. Mulder, Y., Drijver, M. and Kreis, I. 'Case-control study on the association between a cluster of childhood haemotopoietic malignancies and local environmental factors in Aalsmeer, The Netherlands', *Journal of Epidemiology and Community Health*, **48**, 161–165 (1994).
5. Pobel, D. and Viel, J-F. 'Case-control study of leukaemia among young people near La Hague nuclear reprocessing plant: the environmental hypothesis revisited', *British Medical Journal*, **314**, 101–106 (1997).
6. Alexander, F., Wray, N., Boyle, P., Bring, J., Coebergh, J., Draper, G., Levi, F., McKinney, P., Michaelis, J., Peris-Bonet, R., Petridou, E., Pukkala, E., Storm, H., Terracini, B. and Vatten, L. 'Clustering of childhood leukaemia: a European study in progress', *Journal of Epidemiology and Biostatistics*, **1**, 13–24 (1996).
7. Alexander, F. E., Boyle, P., Carli, P-M., Coebergh, J., Draper, G. J., Ekbom, A., Levi, F., McKinney, P., McWhirter, W., Michaelis, J., Peris-Bonet, R., Petridou, E., Pompe-Krin, V., Plěsko, I., Pukkala, E., Rahu, M., Storm, H., Terracini, B., Vatten, L. and Wray, N. R. 'Spatial clustering of childhood

leukaemia: summary results from the EUROCLUS project', *British Journal of Cancer*, **77**, 818–824 (1998).
8. Alexander, F. E., Boyle, P., Carli, P-M., Coebergh, J. W., Draper, G. J., Ekbom, A., Levi, F., McKinney, P. A., McWhirter, W., Magnani, C., Michaelis, J., Olsen, H., Peris-Bonet, R., Petridou, E., Pukkala, E. and Vatten, L. on behalf of the EUROCLUS project. 'Spatial and temporal patterns in childhood leukaemia: further evidence of an infectious origin', *British Journal of Cancer*, **77**, 812–817 (1998).
9. Craft, A. W., Openshaw, S. and Birch, J. 'Childhood cancer in the Northern Region, 1968–1982: incidence in small geographical areas', *Journal of Epidemiology and Community Health*, **39**, 53–57 (1985).
10. Gardner, M. 'Review of reported increases of childhood cancer rates in the vicinity of nuclear installations in the UK', *Journal of the Royal Statistical Society*, *Series A*, **152**, 307–326 (1989).
11. Clayton, D. and Kaldor, J. 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics*, **43**, 671–668 (1987).
12. Clayton, D. and Bernardinelli, L. 'Bayesian methods for mapping disease risk', *in* Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), Geographical and Environmental Epidemiology: Methods for Small area Studies, Oxford Medical Publications, 1992, Chapter 18, pp. 205–220.
13. Potthoff, R. and Whittinghill, M. 'Testing for homogeneity. II. The Poisson distribution', *Biometrika*, **53**, 183–191 (1966).
14. Breslow, N. E. 'Extra-Poisson variation in log-linear models', *Applied Statistics*, **33**, 38–44 (1984).
15. Dean, C. and Lawless, J. F. 'Tests for detecting overdispersion in Poisson regression models', *Journal of the American Statistics Association*, **84**, 467–472 (1989).
16. Muirhead, C. R. 'Childhood leukaemia in metropolitan regions in the United States: a possible relation to population density?', *Cancer Causes and Control*, **6**, 383–388 (1995).
17. Muirhead, C. R. and Butland, B. K. 'Testing for over-dispersion using an adapted form of the Potthoff-Whittinghill method', *in* Alexander, F.E. and Boyle, P. (eds.), *Statistical Methods of Investigating Localised Clustering of Disease*, International Agency for Cancer, Lyon, 1996, pp. 40–52.
18. Collings, B. J. and Margolin, B. H. 'Testing goodness of fit for the Poisson assumption when observations are not identically distributed', *Journal of the American Statistics Association*, **80**, 411–418 (1985).
19. Openshaw, S., Charlton, M., Craft, A. and Birch, J. 'Investigation of leukaemia clusters by the use of the geographical analysis machine', *Lancet*, **i**, 272–273 (1988).
20. Potthoff, R. and Whittinghill, M. 'Testing for homogeneity. I. The binomial and multinomial distributions', *Biometrika*, **53**, 167–182 (1966).
21. Cliff, A. and Ord, J. *Spatial Autocorrelation*, Pion Ltd, London, 1981.
22. Centor, R. 'The use of ROC curves and their analyses', *Medical Decision Making*, **11**, 102–106 (1991).

