Comparison of the Entropy Technique with Two Other Techniques for Detecting Disease Clustering Using Data from Children with High Blood Lead Levels

Joel B. Swartz,1,* Stephen J. Rothenberg,23 Senait Teklehaimanot,1 and Fuad Khan4

The entropy technique was compared with two other case-control techniques for detecting disease clustering using data on blood lead levels of children who were patients at the King/Drew Medical Center in South-Central Los Angeles in 1991 to 1994. The other two methods are the nearest neighbor technique (NNT) and Moran’s IPOP technique, a variation of Moran’s I test, in which rates are adjusted for population size. Four different blood lead levels (15 µg/dl, 20 µg/dl, 30 µg/dl, 35 µg/dl) were used as cutoff levels to designate cases. Persons with blood lead levels greater than or equal to the cutoff level were designated as cases. The authors found significant clustering for all four cutoff levels using the entropy method, and for the first three cutoff levels using the NNT. They found significant clustering with Moran’s IPOP for some scales for two of the cutoff levels. While performance of the entropy technique and the NNT were independent of scale, that of Moran’s IPOP was highly scale-dependent. Am J Epidemiol 1999;149:750–60.

disease clustering; entropy; environmental health; lead

The purpose of this paper is to compare the performance of the entropy technique with two other cluster detection techniques used with a case-control experimental design: the nearest neighbor technique (NNT) and Moran’s IPOP technique, a variation of Moran’s I test, in which rates are adjusted for population size. The three techniques have been applied to a data set consisting of blood lead levels of children who were patients at the King/Drew Medical Center in South-Central Los Angeles over the period 1991–1994 (1). This data set was chosen because exploratory analyses indicated substantial geographic differences in average blood lead levels (1).

DISEASE CLUSTERING AND DETECTION TECHNIQUES

Special mathematical techniques are needed to detect clusters of disease because of spatial inhomogeneities in population density. A disease cluster is an area with an especially elevated disease rate. Without the use of clustering techniques, diseases might appear to cluster specifically because many cases are located in regions of high population density. All statistical disease clustering techniques test the null hypothesis, that the disease rate is constant throughout the entire space under study. The output they yield is a p value, which is the probability of achieving the observed value of the clustering statistic, or a value indicating clustering greater than that observed under the null hypothesis.

There are several types of spatial clusters, and some techniques are most appropriate for certain types of clusters. Walter (2) reviews a variety of types of clusters, and some techniques appropriate for each. These cluster types include: hot-spot cluster type (single area of high disease incidence), directional gradient in disease incidence type and general positive spatial autocorrelation pattern type (i.e., nearby areas have related rates, incidence rate correlated with external variable such as urbanization or socioeconomic level). Another common pattern is that of several spatially separated focused clusters in the space of interest. Other cluster-
ing methods are omnibus techniques, i.e., they are able to detect any type of spatial clustering.

Many disease clustering techniques use tests based on the number of adjacent regions with similar disease rates (3). These techniques require the division of a geographic area into a number of regions, and the categorization of these regions by disease rate. These techniques tend to have low power because they discard much of the information in the data. Instead of using actual locations of cases, or actual disease rates in certain areas, these techniques make use only of a category of disease rate in each area (e.g., the rate of 15 per 100,000 per year might be replaced by the category “medium rate”).

In general, clustering techniques that are used with data in a case-control format are more powerful than adjacency based methods. Case-control methods are commonly used in epidemiologic studies of rare diseases to reduce the required sample size. With these techniques, all subjects in the study are either persons with the disease (cases) or persons without the disease (controls), and the number of controls is comparable (within a factor of three) with the number of cases.

The most widely used clustering technique for case-control studies is the NNT of Cuzick and Edwards (4). The NNT is based on the principle that if there is disease clustering the subjects geographically closest to cases are more likely to be cases than would be expected under the null hypothesis. The test statistic, \( T_k \), is the sum over all cases of the number of each case’s \( k \) nearest neighbors that are also cases.

Swartz (5) has developed the entropy method for use in case-control studies. In this technique, the space under study is divided into a number of cells, and the number of cases and controls in each cell is counted. Then the number of possible arrangements of the total number of cases and controls in the different cells, consistent with the total in each cell, is computed. The log of the number of possible arrangements is proportional to the entropy. If the disease is highly clustered, then the number of possible arrangements will be small. However, if the disease is distributed homogeneously, then there will be many possible arrangements. In general, high entropy is associated with a high degree of disorder, and low entropy with a high degree of order. This concept of entropy is similar to the concept of entropy in thermodynamics, statistical mechanics, and information theory (6). The basic entropy technique is explained in Appendix 1.

The entropy technique was developed to overcome the following weaknesses of the NNT:

1. In the NNT, some cases will be located outside the region of clustering (i.e., outside of the region where the disease rate is very high). These cases will have fewer than the expected number of cases as nearest neighbors. This phenomenon will result in the cancellation of some of the clustering effect. The entropy method measures only inhomogeneity. Therefore, cells with more than the expected number of controls will add to the overall measure of clustering.

2. The NNT is sensitive to the shape of the cluster. If the cluster is circularly shaped, \( T \) is likely to be very high. On the other hand, if the cluster is pipe-shaped, the cases will still have many controls as nearest neighbors, and \( T \) will be lower than for the circular-shaped cluster. Therefore, two clusters with the same degree of clustering, but different shapes, may have very different values of \( T \).

3. The NNT is sensitive to edge effects. If a cluster is located near the edge of the space, the cases may have many controls as nearest neighbors, and \( T \) will be low. Thus, two clusters with the same degree of clustering, but located with differing proximity to the space boundaries, may have very different \( T \) values.

4. The NNT summary statistic does not accurately measure the improbability of a spatial arrangement. It is possible for two arrangements with substantially different probabilities to have the same value of \( T \) (5). Because hypothesis tests in statistics are based on determining the improbability of a given result under the null hypothesis, it is desirable to use a test statistic which accurately measures the probability of a situation given the null hypothesis.

In a series of tests with synthetic data, the entropy method was found to be more powerful than the NNT for several situations (5). Specifically, the entropy method was more powerful than the NNT in situations in which there were two or more separate clusters in the space, or in which the clusters were situated at the edge of the space. In some instances, the difference in power exceeded 20 percent.

A third technique, which is suitable for case-control data, is Moran’s IPOP method (7, 8), in which rates are adjusted for population size. Previously, this method has been applied to data on disease rates. In this method, the space is divided into regions and the disease rate is computed in each region. Then a weighted correlation function, IPOP, is computed. IPOP measures the correlation between rates in different regions. \( I \) is heavily weighted by the inverse of the distance between regions. The formula for IPOP also includes weighted terms for the square of the difference between the observed and expected number of cases in each region. In the original version of Moran’s \( I \) test,
there was no adjustment for population size. Oden (8) developed the IPOP statistic which is similar to the I test except that, with IPOP, regions are weighted by population size. We used this form of Moron’s test so that areas with small numbers of people are not over-weighted in the clustering calculations. Without this correction, a small number of cases or controls in an area with a low population density might cause an apparent cluster when there is none, or cause a real cluster to go undetected.

Therefore, two phenomena can elevate the value of the IPOP statistic: 1) high correlations in rates between nearby regions, and 2) large differences between the observed and expected number of cases in any region. IPOP increases with the degree of clustering. Although the IPOP method was originally designed for use with population disease rates, it can be adapted readily for use with data in a case-control format. In this situation, the probability of being a case in a region replaces the disease rate for the region.

**Population screening for high lead levels**

Lead poisoning has long been a serious problem in the United States, and it has a substantially higher prevalence among children who live in inner city areas (9). Previous studies have shown that lead has a detrimental effect on the neurologic development of young children, and of children who had been exposed in utero (10). Blood lead levels of children above 10 μg/dl are considered dangerous (11), while those above 20 μg/dl are considered very dangerous, and require immediate attention (11).

The catchment area of the King/Drew Medical Center includes many older dwellings with lead paint, and at least one industrial site known to emit large quantities of lead. A previous study (1) suggested some spatial clustering of high lead levels, because there were substantial differences in average lead levels between zip codes. There was high correlation between average income and geographic mean blood lead, grouped by zip code (r = -0.7) (1). Although there was reason to anticipate clustering of children with high lead levels, the pattern of clustering was unknown. The three cluster detection techniques were applied to the childhood blood lead data set to determine if there was such clustering.

**METHODS**

The data for this study were measurements of blood lead taken from all 3,679 children screened at the King/Drew Medical Center over the period 1991–1994. The results of only the first test were used for children who were tested more than once.

The NNT, entropy, and Moran’s IPOP techniques were applied for the following four blood lead cutoff values: 15 μg/dl, 20 μg/dl, 30 μg/dl, 35 μg/dl. In each instance, a subject with a blood lead value equal to or greater than the cutoff value was considered a case. All other subjects were placed in the pool from which controls were selected. Because the cutoff level defined whether an individual was diseased or not it was appropriate to include all individuals below the cutoff level as potential controls. The actual number of potential controls who had high blood lead levels was very small, e.g., fewer than 4 percent had blood lead levels above 15 μg/dl. The inclusion of a few subjects with lead levels above 15 μg/dl in the control pool is unlikely to affect the outcome. For each case, two controls matched on race/ethnicity, sex, and age category were randomly selected without replacement from the pool of all potential controls within the same demographic category. The age categories were 0–4, 5–9, and 10–18 years. Ninety-eight percent of the subjects were either African-American or Hispanic. Because there is no matching by region, there is no risk that the process of selecting controls will either create an artificial cluster or cause a real cluster to go undetected. The calculations for both Moran’s IPOP and the NNT were performed using STAT Version 2.2 (Biomedware, Ann Arbor, Michigan). For both the entropy method and Moran’s IPOP, the space was divided into square cells using grids with the following dimensions: 5 x 5, 7 x 7, 9 x 9. STAT Version 2.02 reports p values for all numbers of nearest neighbors from 1 to 10. It also reports two summary significance values taking into account all 10 results combined, and correcting for multiple comparisons. In the Discussion, we consider the Simes (12) overall evaluation statistic for the NNT because its power and alpha level are most in accord with results from individual significance tests. We calculated the probability of certain individual values for specific numbers of nearest neighbors. Because the entropy method and Moran’s IPOP both use the same grid sizes, and because the grid size may affect the results, we have discussed the results for the three grid sizes for these methods. Additionally, because the NNT method does not use gridding, it was not possible to make exact comparisons with the other two methods on a grid size by grid size basis. In almost all instances, the scale factor, i.e., the grid size or number of nearest neighbors used in the NNT computation was not a factor in the performance of either the entropy technique or the NNT method.

For these data, we used a variant of the entropy method in which a likelihood function for the arrangement of the cases and controls in the cells of the area
investigated was used as the test statistic. This method was used because the number of controls was twice the number of cases. In Swartz (5), it was shown that in situations in which the numbers of cases and controls are far from equal, the state of highest entropy is not necessarily the least clustered state. However, it has also been demonstrated (5) that the likelihood, when calculated appropriately, decreases monotonically with increasing clustering.

This variation of the entropy technique could also be derived directly from likelihood or probability considerations, but it is also a natural extension of the entropy method. It involves the number of possible arrangements of cases and controls with a weighting for the probability of being a case or a control. A detailed explanation of this modified entropy technique is given in Appendix 2.

**RESULTS**

Significant clusterings of high blood lead levels were found. In each situation, there was clustering in several geographically distinct areas. Tables 1 and 2 show the longitude and latitude of the centroids of the cells with high proportions of cases at the cutoffs of 15 μg/dl, and 20 μg/dl for the 5 × 5 and 7 × 7 grids, respectively. For each cell, the probability of being a case, and the total number of subjects is shown. By design, one-third of the subjects were cases. Figures 1 and 2 show the number of cases (upper number) and controls (lower number) in each cell. In all of the maps shown here, the origin of column and row number is in the lower right corner.

**TABLE 1.** Geographic regions with high proportions of cases with cutoff value at 20 and 15 μg/dl—children with high blood lead levels who were patients at King/Drew Medical Center, Los Angeles, California, 1991–1994: using 5 × 5 grid, background case probability is 0.33

<table>
<thead>
<tr>
<th>Longitude</th>
<th>Latitude</th>
<th>Column no.</th>
<th>Row no.</th>
<th>Probability of being case</th>
<th>Total subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>-118.315</td>
<td>33.952</td>
<td>5</td>
<td>3</td>
<td>0.67</td>
<td>3</td>
</tr>
<tr>
<td>-118.189</td>
<td>33.924</td>
<td>2</td>
<td>2</td>
<td>0.50</td>
<td>14</td>
</tr>
<tr>
<td>-118.231</td>
<td>34.008</td>
<td>3</td>
<td>4</td>
<td>0.80</td>
<td>2</td>
</tr>
<tr>
<td>-118.147</td>
<td>33.896</td>
<td>1</td>
<td>1</td>
<td>0.50</td>
<td>4</td>
</tr>
<tr>
<td>-118.147</td>
<td>33.924</td>
<td>1</td>
<td>2</td>
<td>0.82</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Longitude</th>
<th>Latitude</th>
<th>Column no.</th>
<th>Row no.</th>
<th>Probability of being case</th>
<th>Total subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>-118.273</td>
<td>33.924</td>
<td>4</td>
<td>2</td>
<td>0.47</td>
<td>19</td>
</tr>
<tr>
<td>-118.231</td>
<td>33.980</td>
<td>3</td>
<td>4</td>
<td>0.75</td>
<td>8</td>
</tr>
<tr>
<td>-118.189</td>
<td>33.896</td>
<td>2</td>
<td>1</td>
<td>0.80</td>
<td>5</td>
</tr>
<tr>
<td>-118.189</td>
<td>34.008</td>
<td>2</td>
<td>5</td>
<td>0.67</td>
<td>3</td>
</tr>
<tr>
<td>-118.147</td>
<td>33.924</td>
<td>1</td>
<td>2</td>
<td>0.83</td>
<td>18</td>
</tr>
</tbody>
</table>

**TABLE 2.** Geographic regions with high proportions of cases with cutoff value at 20 and 15 μg/dl—children with high blood lead levels who were patients at King/Drew Medical Center, Los Angeles, California, 1991–1994: using 7 × 7 grid, background case probability is 0.33

<table>
<thead>
<tr>
<th>Longitude</th>
<th>Latitude</th>
<th>Column no.</th>
<th>Row no.</th>
<th>Probability of being case</th>
<th>Total subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>-118.291</td>
<td>33.852</td>
<td>6</td>
<td>4</td>
<td>0.40</td>
<td>5</td>
</tr>
<tr>
<td>-118.231</td>
<td>33.972</td>
<td>4</td>
<td>5</td>
<td>0.80</td>
<td>5</td>
</tr>
<tr>
<td>-118.201</td>
<td>33.892</td>
<td>3</td>
<td>2</td>
<td>0.40</td>
<td>5</td>
</tr>
<tr>
<td>-118.201</td>
<td>33.932</td>
<td>3</td>
<td>4</td>
<td>0.67</td>
<td>6</td>
</tr>
<tr>
<td>-118.201</td>
<td>33.932</td>
<td>3</td>
<td>4</td>
<td>0.57</td>
<td>7</td>
</tr>
<tr>
<td>-118.171</td>
<td>33.932</td>
<td>2</td>
<td>3</td>
<td>0.60</td>
<td>5</td>
</tr>
<tr>
<td>-118.141</td>
<td>33.932</td>
<td>1</td>
<td>3</td>
<td>0.50</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Longitude</th>
<th>Latitude</th>
<th>Column no.</th>
<th>Row no.</th>
<th>Probability of being case</th>
<th>Total subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>-118.261</td>
<td>33.912</td>
<td>5</td>
<td>2</td>
<td>0.40</td>
<td>10</td>
</tr>
<tr>
<td>-118.261</td>
<td>33.952</td>
<td>5</td>
<td>4</td>
<td>0.52</td>
<td>27</td>
</tr>
<tr>
<td>-118.201</td>
<td>33.952</td>
<td>3</td>
<td>4</td>
<td>0.53</td>
<td>15</td>
</tr>
<tr>
<td>-118.201</td>
<td>34.012</td>
<td>3</td>
<td>7</td>
<td>0.67</td>
<td>3</td>
</tr>
<tr>
<td>-118.171</td>
<td>33.892</td>
<td>2</td>
<td>1</td>
<td>0.60</td>
<td>5</td>
</tr>
<tr>
<td>-118.171</td>
<td>33.932</td>
<td>2</td>
<td>3</td>
<td>0.62</td>
<td>13</td>
</tr>
<tr>
<td>-118.141</td>
<td>33.932</td>
<td>3</td>
<td>3</td>
<td>0.62</td>
<td>13</td>
</tr>
</tbody>
</table>

In order to interpret the results, it is useful to use the concept of the odds ratio. The odds ratio (OR) for a disease for any factor is defined as:

\[
OR = \frac{pcaseWF \times (pconWOF)}{(pconWF \times pcaseWOF)}
\]

where pcaseWF is the probability of being a case with the factor, pconWOF the probability of being a control without the factor, pconWF the probability of being a control with the factor, and pcaseWOF is the probability of being a case without the factor.

Here the factor of interest is residence within a given cell. The odds ratio is a measure of the relative probability of having the disease if one has the factor, i.e., living in a particular cell on the grid.

To help interpret the results, note that for areas in which 60 percent of the subjects are cases, the odds ratio is 3, and for areas in which 80 percent of the subjects are cases the odds ratio is 8.

The odds ratios for the cutoff of 15 μg/dl and 20 μg/dl are shown on the maps in figures 3 and 4. The circular symbols are proportional in size to the case-control odds ratio in the particular cell. The asterisks indicate regions with less than the expected number of cases, or regions with no cases.

The significance of the clustering found using the different methods for various blood lead cutoff values is shown in table 3.

FIGURE 3. Geographic distribution of high blood lead levels in children in catchment area of King/Drew Medical Center, South-Central Los Angeles, California, 1991–1994. Odds ratios by cell for 5 x 5 grid, 15μg/dl cutoff. Large circular symbols indicate regions with high degree of clustering. Small circular symbols indicate regions with moderate degree of clustering. Asterisks indicate regions with fewer than the expected number of cases or with no cases. Frwy, freeway.

FIGURE 4. Geographic distribution of high blood lead levels in children in catchment area of King/Drew Medical Center, South-Central Los Angeles, California, 1991–1994. Odds ratios by cell for 5 x 5 grid, 20μg/dl cutoff. Large circular symbols indicate regions with high degree of clustering. Small circular symbols indicate regions with moderate degree of clustering. Asterisks indicate regions with fewer than the expected number of cases or with no cases. Frwy, freeway.
For the NNT, one summary \( p \) value was computed by calculating the \( p \) values for all values of nearest neighbors from 1 to 10, and then applying the Simes correction for multiple comparisons (12). For the entropy method and for Moran’s IPOP test, separate \( p \) values were calculated for grids with the following dimensions: 5 \( \times \) 5, 7 \( \times \) 7, and 9 \( \times \) 9.

For the cutoff of 15 \( \mu g/dl \), there were 134 cases. We found significant clustering with the entropy method for all three grid sizes (\( p \leq 0.015 \)), for Moran’s IPOP with all three grid sizes (\( p \leq 0.026 \)) and for the NNT (\( p = 0.0118 \)). For both the entropy and the Moran’s IPOP techniques, we found the lowest \( p \) values using the 9 \( \times \) 9 grid.

For the cutoff of 20 \( \mu g/dl \), there were 53 cases. We found highly significant clustering with the entropy method for all three grid sizes (\( p \leq 0.001 \)), and also the NNT (\( p = 0.0038 \)). We did not find significant clustering with Moran’s IPOP for any grid size (\( p \geq 0.071 \)).

For the cutoff of 30 \( \mu g/dl \), there were 15 cases. We found highly significant clustering with the entropy technique for all three grid sizes (\( p \leq 0.001 \)), and also with the NNT (\( p \leq 0.0001 \)). We found significant clustering with Moran’s IPOP for the 7 \( \times \) 7 grid (\( p = 0.0020 \)), and the 9 \( \times \) 9 grid (\( p = 0.0059 \)), but not for the 5 \( \times \) 5 grid (\( p = 0.16 \)).

Not shown in table 3 is that there were nine cases with a cutoff of 35 \( \mu g/dl \). We found significant clustering with the entropy method for all three grid sizes (\( p \leq 0.004 \)). We did not find significant clustering with the NNT (\( p = 0.0538 \)), or with Moran’s IPOP for any grid size (\( p \geq 0.36 \)).

**DISCUSSION**

The entropy method and the NNT showed similar sensitivity for three of the four situations investigated. In the situation with the fewest cases (cutoff = 35 \( \mu g/dl \), 9 cases), we found highly significant clustering for all three grid sizes using the entropy method, but did not find significant results using the NNT. This is an instance for which the entropy method was more sensitive than the NNT.

Significant clustering was found with Moran’s IPOP for the cutoff of 15 \( \mu g/dl \) (134 cases) using all three grid sizes. For the cutoff 20 \( \mu g/dl \), this test did not show significant clustering at any grid size, while for the 30 \( \mu g/dl \) cutoff, there was significant clustering for the 7 \( \times \) 7 and 9 \( \times \) 9 grid sizes. Significant clustering was not detected with the Moran’s IPOP technique for any grid size when the cutoff was at 35 \( \mu g/dl \) (nine cases).

Overall, the NNT and the entropy technique were far more sensitive than Moran’s IPOP. Additionally, the performance of the entropy test was not strongly affected by the grid size, while the \( p \) values varied dramatically with grid size for Moran’s IPOP. There was a small nonsignificant trend of declining \( p \) values with increasing grid size for the entropy technique.

Because the \( p \) values for the entropy method are low, reflecting a statistic in the upper tail, large changes in the statistic will produce small changes in the \( p \) value.
It is possible, that for a less clustered situation, the entropy method would be more sensitive to grid size.

One reason for the general weakness of Moran’s IPOP for this study is that for most cutoff values there were several separate clusters which were spatially separated by areas with low or average rates of children with high blood lead levels. The use of Moran’s IPOP technique is based on the premise that, in situations of clustering, correlations in case probability decline as the distance between regions increases. Because there were several clusters separated by areas with no clustering, this premise is not generally true in this study.

An apparent anomaly is that for the NNT and Moran’s IPOP, the p values declined (the significance of the clustering increased) as the cutoff was extended from 20 μg/dl (53 cases) to 30 μg/dl (15 cases). The explanation is that the dispersion of clusters generally reduces the power of Moran’s IPOP, and sometimes reduces the power of the NNT. Therefore, as the number of cases was reduced from 53 to 15, the number of separate clusters declined, and the significance of the clustering increased. This situation is described in Appendix 3.

Hence, it is reasonable to use Moran’s IPOP for spatial clustering with case-control data only in situations in which it is expected that there is a single hot-spot or focused cluster. On the other hand, both the NNT and the entropy techniques can be used when there are a number of spatially separated clusters, although the NNT may have relatively low power for some situations with dispersed clusters.

In this study and in previous studies (5), both the entropy method and the NNT have been shown to detect a wide variety of cluster types. Therefore, they can be used as omnibus clustering methods when nothing is known about the cluster type, before the study is performed.

As noted, tables 1 and 2 indicate the five regions with the highest case probability for the cutoff values of 15 μg/dl and 20 μg/dl for the 5 × 5 and 7 × 7 grid sizes, respectively. The majority of the highest regions with the 15 μg/dl cutoff are different from the highest regions with the 20 μg/dl cutoff. This is likely because the factors which produce the blood lead levels in the vicinity of 15 μg/dl are different from the factors which produce very high blood lead levels, i.e., over 20 μg/dl.

CONCLUSIONS

1. There is significant non-contiguous clustering of children with high blood levels in the catchment area of the King/Drew Medical Center. Based on the proportions of cases and controls in different cells, the clustering apparently occurs in several separate geographic areas.

2. The entropy method and the NNT had similar sensitivities in detecting clusters in most of the situations examined in this study. However, for the situation with only nine cases, the entropy method was more sensitive than the NNT.

3. Moran’s IPOP, in general, is less sensitive than the other two methods for the type of clustering examined in this study. In large part, its weakness is due to its inability to detect clustering when there are two or more dispersed clusters. It should be used when hot-spot clustering is anticipated.

4. The entropy and the NNT perform well when there are several dispersed clusters. They can be used as omnibus clustering methods when nothing is known before the study concerning the type of clustering.

5. Moran’s IPOP is highly sensitive to grid size, whereas the entropy method is relatively insensitive to grid size within a reasonable size range.

6. The entropy method can indicate regions of clustering through the p values and odds ratios, while the NNT does not.

ACKNOWLEDGMENTS

The work in this paper was supported by the Research Centers in Minority Institutions (Epidemiology/Statistical Research Group) Grant (NCCR/NIH/RCMI/G12 RR 03026) from the National Institute of Health and by the Development of the Entropy Method for Cluster Detection Grant (1 R03 CA71614-01A) from the National Cancer Institute.

The authors thank Phyllis Bryant for help with manuscript preparation.

REFERENCES


APPENDIX 1

Description of Basic Entropy Method

The idea of the entropy method is that if two types of objects are spread homogeneously through a space there will be many possible ways for the objects to be arranged in the space. That is, any object could be placed almost anywhere in the space. This corresponds to the unclustered situation, to a highly ordered situation. On the other hand, if there is clustering, one type of object will tend to be in one region in the space, and the other type in another region. There will be fewer possible ways to arrange the objects in the space.

The entropy is calculated by dividing the space into a number of cells. Consider the simple situation illustrated in figures 5 and 6. There are two cases (indicated by X), two controls (indicated by O), and the space is divided into two cells. The two cases are distinguishable from each other as are the two controls.

In figure 5, the two controls are in the first cell, and the two cases in the second cell. This is the maximally clustered situation. There is only one possible way to arrange the cases and controls to make up this situation. Both controls must be placed in the first cell, and both cases in the second. The entropy is log 1 = 0.

In figure 6, each cell has one case and one control. This is the unclustered situation. Now the number of ways to make up this arrangement is as follows: There are two possible ways to select the case and two possible ways to select the control for the first cell. Consequently, there are $2 \times 2 = 4$ ways to arrange this situation. Thus, the entropy is log 4 = 1.38. Therefore, the entropy in the unclustered or homogeneous situation is higher than the entropy of the clustered situation. This is the principle of the entropy method.

APPENDIX 2

Description of the Entropy Method Used in This Study

This method is a likelihood-based method which is conceptually and mathematically close to the original entropy method. It is possible to derive it from other concepts. Its advantage over the original entropy method is that its efficacy is independent of the ratio of cases to controls.

The basic idea is that if the disease is distributed homogeneously then the underlying probability of being a case in each cell will be the same as the probability of being a case in the entire space. Also the proportion of cases in a cell will actually be close to the proportion of cases in the entire space, with the actual distribution being binomial. If the disease is distributed inhomogeneously, then the expected proportion of cases in some cells will be lower than that expected in the overall space, and higher in others. The actual propor-

![Figure 5](https://example.com/figure5.png)

**FIGURE 5.** Calculation of entropy for situation with 2 controls, 2 cases, 2 cells. The two cases are in the first cell, and two controls are in the second cell. Entropy in clustered situation. X indicates cases. O indicates controls. Entropy = 0.

![Figure 6](https://example.com/figure6.png)

**FIGURE 6.** Calculation of entropy for situation with 2 controls, 2 cases, and 2 cells. Each cell with one case and one control. Entropy in unclustered or homogeneous situation. X indicates cases. O indicates controls. Entropy = 1.38.
tions will also be different in different parts of the space.

The $E$ function for any particular cell is the probability of obtaining the observed proportion of cases in that cell given the overall number of cases and controls in the whole space. It is the product of $E_i$ for each cell. $E_i$ is the likelihood of obtaining the arrangement in cell $i$ given that it can be selected from all the available cases and controls. The formula is as follows:

$$E_i = p \times \times \text{NCASE}_i \times (1 - p) \times \times \text{NCON}_i \times \frac{(\text{NTOT}!)}{(\text{NCASE}_i!)(\text{NCON}_i!)}$$

where $E_i$ is the contribution to the $E$ function from the $i$th cell, $p$ is the overall probability of being a case in the space, NCASE$_i$ is the number of cases in cell $i$, NCON$_i$ is the number of controls in cell $i$, and NTOT is the total number of subjects in the cell $i$. See, for example, Mendenhall et al. (13), section 3.4.

The total $E$ function is the product of the $E_i$ from each cell. The overall determination of clustering is made from the total $E$ function or the log of $E$ which is equivalent.

This choice of the $E$ function was reached in both theoretical and empirical grounds. It is the same as the likelihood function for a binomial distribution assuming the cases and controls were sampled with replacement from the total of cases and controls. The final form of the $E$ function was reached after experimental testing of several candidates.

**Example of calculation of $E$ function**

Suppose we have three cells with cases and controls as follows:

<table>
<thead>
<tr>
<th>Cell</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Overall, there are nine controls and four cases. So $p$ is $4/13 = 0.31$

$$E_1 = 0.31^1 \times 0.69^2 \times 3!/1! \times 2! = 0.443$$

$$E_2 = 0.31^2 \times 0.69^2 \times 4!/2! \times 2! = 0.275$$

$$E_3 = 0.31^1 \times 0.69^5 \times 6!/1! \times 5! = 0.290$$

$$E = E_1 \times E_2 \times E_3 = 0.0353$$

$\ln E = -3.34$

The actual calculations for the method proceeded as follows:

1. Calculate $E$ for the entire space as above.
2. After the calculation with the actual data was performed, the space was randomized 1,000 times and the likelihood calculated for each randomization. By randomization, we mean that the X and Y coordinates of the cases and controls are kept fixed, but the assignment as a case or control was randomly determined with the constraint that the total number of cases and controls equaled the total number of cases and controls in the data set.
3. The $p$ value for the observed situation was determined using the likelihood for the randomized situations. The $p$ value is the probability of obtaining the observed likelihood value or a more extreme value if the null hypothesis (no clustering) is true. For example, suppose the observed likelihood were $-3.23$. Suppose that of the 1,000 likelihoods calculated from the randomizations, $-3.23$ corresponded to number 956, i.e., the 44th from the bottom. Then the probability of obtaining the observed likelihood or a more extreme value would be approximately 0.044.

Therefore the estimate of the $p$ value is 0.044.

**APPENDIX 3**

Figures 7 and 8 (see next page) illustrate two situations with the same degree of clustering as measured by the entropy technique and the NNT, but in which the significance of clustering computed with Moran's IPOP is substantially different. In both situations, there are two regions of clustering. However, in the situation shown in figure 7 the two regions of clustering are widely separated, while in that shown in figure 8 the regions of clustering are adjacent. For the NNT, the overall $p$ value for situation A and B is both 0.001. For Moran's IPOP, the $p$ value for situation A is 0.31, while for situation B it is 0.20. For the entropy method, the $p$ values are the same in each instance. The $p$ value is 0.003.

This illustrates that the sensitivities of the entropy method and the NNT method do not vary as the distance between two separate regions of clustering varies, but that the sensitivity of Moran's IPOP can be greatly affected.
FIGURE 7. Comparison of significance of clustering between situation with adjacent clusters and situation with widely dispersed clusters using NNT, entropy, and Moran's IPOP. Two adjacent clusters. X indicates cases. O indicates controls. $p$ values: entropy, 0.003; NNT, 0.001; Moran's IPOP, 0.20.

FIGURE 8. Comparison of significance of clustering between situation with adjacent clusters and situation with widely dispersed clusters using NNT, entropy, and Moran's IPOP. Two clusters separated by large distance. X indicates cases. O indicates controls. $p$ values: entropy, 0.003; NNT, 0.001; Moran's IPOP, 0.20.